Memory and Representativeness

Pedro Bordalo Katherine Coffman Nicola Gennaioli
Frederik Schwerter Andrei Shleifer*

March 5, 2019

Abstract

We explore the idea that judgment by representativeness reflects the workings of episodic memory, especially interference. In a new laboratory experiment on cued recall, participants are shown two groups of images with different distributions of colors. 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) for a fixed set of images, different cues for the same objective distribution entail different interference patterns and different probabilistic assessments. Selective retrieval and interference may offer a foundation for the representativeness heuristic, but more generally for understanding the formation of probability judgments from experienced statistical associations.

*Bordalo: University of Oxford; pedro.bordalo@sbs.ox.ac.uk. Coffmann: 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, Sam Gershman, Thomas Graeber, Joshua Schwartzstein, and seminar participants at Harvard University and MIT Sloan 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 a range of systematic errors (see Benjamin 2018 for a review). Kahneman and Tversky (KT 1972) argue that these intuitive judgments reflect the differential accessibility of information in the mind: events or instances that are more accessible are perceived as being more likely. "To understand intuition, then, we must understand why some thoughts are accessible and others are not" (Kahneman 2003).

A central body of evidence focuses on the representativeness heuristic, the tendency to judge as likely events that are merely representative. As KT define it when discussing base-rate neglect and the conjunction fallacy, representativeness captures "the degree to which [an event] is similar in essential characteristics to its parent population" (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 (KT 1974). Most subjects state that the introverted man is more likely to be a librarian than a farmer, even though there are vastly more male farmers than librarians. The thought that an introverted man is a librarian is highly accessible because the trait "introvert" makes him more similar to a typical librarian than to a typical farmer. In this view, representative events are highly accessible and thus play an outsized role in probability judgments. In fact the ranking of items in terms of probability is often aligned with that in terms of similarity in such experiments.

This account does not explain why features that are perceived as representative, or similar, are often so unlikely that they cause judgments to be incorrect. Why, among introverted men, do we think about the very unlikely librarian instead of the more likely farmer? In a different example, why do we think it is more representative of Hollywood actresses to be "more than 4-times divorced" rather than to "vote Democrat", which is much more likely (KT 1984)? One possibility is that stereotypes of introverted male librarians and oft-divorced Hollywood actresses are spread through the media and personal interactions. This semantic association and repetition of unlikely traits becomes a source of similarity and accessibility in memory. Bhatia (2017) presents evidence on natural language consistent with this mechanism. While this mechanism certainly contributes to biases in social contexts, it does not explain the source of these biases (why does natural language focus on these specific traits), and why biases are widespread even in more abstract judgments.

In this paper we propose that the accessibility of unlikely instances reflects how we represent statistical associations among multiple features, and especially how we retrieve them from memory when making judgments. In our view, even if one's personal experience is unbiased, selective retrieval can lead to systematic overweighting of cer-

tain features. Our approach builds on a large literature on memory stressing that recall is selective (see Kahana 2012 for a review). First, a cue such as "an introverted man" or "a Hollywood actress" triggers the recall of memory traces that are similar to it, in the specific sense of frequently co-occurring with the cue. Second, and crucially, the effect of frequent co-occurrence is dampened by interference: if a feature co-occurs with multiple cues, it is less likely to be retrieved by any one of these cues. We suggest that this mechanism, illustrated in word-pair recall tasks (Kahana 2012) and by the fan effect (Anderson 1974), plays a key role in generating judgment by representativeness.

Our model seeks to capture the effect of memory and interference in probability judgments. In this model, when judging the probability of an event, say that an introverted man is a librarian, we retrieve memories of introverted men from different occupations and perform our assessment based on the set of instances that come to mind. We assume that this retrieval process is subject to two specific forms of interference. First, interference across types within a group: when thinking of introverted men, the presence of a common occupation, say farmer, interferes with the recall of less frequent ones, say librarian, reducing the latter's assessed probability. Second, and crucially, interference across groups: if an occupation, say farmer, is common across groups then it becomes harder to recall it for a specific cued group, say introverted men. The numerous instances of extroverted farmers interfere with the retrieval of introverted ones. In contrast, because there are fewer librarians among extroverts, then it becomes easier to recall introverted librarians. This boosts the probability attached to the latter.

We show that these two forms of interference yield the view of representativeness in Tversky and Kahneman (1983):"an attribute is more representative of a class if it is very diagnostic; that is, if the relative frequency of this attribute is much higher in that class than in the relevant reference class". Thus, the cue "introverted man" triggers recall of librarians and not farmers, despite the fact that the latter may be more likely, because thinking about farmers brings to mind many instances of non-introverted people. These memories interfere with the recall of introverted farmers. Likewise, "Hollywood actress" cues recall of multiple divorces because other features, such as being a Democrat, are common in many occupations, which interferes with the recall of Democratic actresses.

More broadly, the model is in line with, and provides a formal approach for, the intuition that probabilistic assessments reflect similarity judgments (KT 1972). In fact, our model provides a statistical measure of similarity between two events – a cue and a type – that is shaped by interference across cues and thus depends on the full set of events in memory. Evidence for this non-euclidean nature of similarity judgments has already been provided by Tversky (1977). We return to it in Section 5.

A model built on this idea yields several testable predictions on how biases depend on the statistical associations observable in the data. To test these predictions, spelled

out below, we run a series of experiments. In the baseline study, participants observe a sequence of 25 target images and 25 decoy images. Target images consist of 15 blue numbers and 10 orange numbers. In one treatment, denoted $gray$, decoy images are gray shapes. In the other treatment, denoted $blue$, decoy images are blue words. In each treatment, the 50 images are presented one at a time and in a random order. Participants are then told that a number is drawn at random from the images they saw, and are asked to guess what the likely color of that number is. The key goal is to understand how interference, as generated by the decoy distribution, impacts answers to this and other questions about the target distribution. This setup enables us to abstract from semantic associations or pre-existing associations, and to test directly for the role of interference.

Our model predicts that participants' probabilistic judgments about the target distribution – the likely color of a randomly drawn number – will be interfered with by the color distribution of the decoys. Specifically, when a color, say blue, becomes more common in the decoy distribution, it interferes more with the recall of instances of blue numbers. In contrast, there is no interference for orange, which is absent in the decoy distribution. As a consequence, as the frequency of blue objects increases in the decoy distribution, the frequency of blue numbers estimated by the subjects goes down and that of orange numbers goes up.

We find robust evidence in support of this prediction. In the baseline experiment, when asked to predict the likely color of a randomly drawn number from the images they saw, participants are much more likely to give the correct answer blue when the decoy distribution is gray shapes (65% of the time) than when the decoy distribution is blue words (45%), suggesting that indeed the blue words of the decoy distribution interfere with the recall of the blue numbers from the target distribution. In follow-up experiments, we incorporate orange words into the decoy distribution. As our model predicts, as the number of orange words in the decoy distribution increases and the number of blue words decreases (so that interference for orange numbers intensifies), the likelihood of participants believing that orange is the likely color of a randomly drawn number falls. Across all our experiments, interference generated by the decoy distribution consistently impacts probabilistic judgments about the target distribution.

We next address several potential confounds. First, adding distraction tasks between the observation of images and the assessment of distributions does not change the results, suggesting that our findings reflect recall and not mere access to working memory. Second, differential attention to a particular color of numbers is unlikely to explain the evidence: results do not change if we use other dimensions such as font size. Third, because judgments are based on recall of directly observed distributions, our results cannot be attributed to the tendency to confound conditional probabilities with inverse conditional probabilities (e.g. the cab problem, Kahneman and Tversky 1984).

One important implication of our model is that biases depend on the way an assessment is elicited, because different formulations can cue recall of different items and thus influence interference. To assess this implication, we develop a variation of the baseline experiment in which images differ along three dimensions: content (either word or number), color (either blue or orange) and font size (either small or large). Again, we generate a target distribution: 25 numbers, 15 of which are blue and small in font size and 10 of which are orange and large in font size. Our decoy distribution is 25 blue words in a large font size. In this experiment, color and font size are perfectly correlated among the numbers: all orange numbers are also large in font size, while all blue numbers are also small in font size. This ensures that asking participants about the likelihood of a given color or of a given font size among the numbers is objectively equivalent.

Rather than vary the decoy distribution across participants, in this experiment we simply vary the cue. Some participants are cued to think in terms of color, and asked to guess the likely color of a randomly drawn number. Other participants are cued to think in terms of font size, and asked to guess the likely font size of a randomly drawn number. With our construction, the answer is identical for both cues: a number is equally likely to be blue and small in font size. However, we hypothesize that the pattern of interference in recall is different. When cued to think in terms of color, the blue words in the decoy distribution will interfere with the recall of blue (small) numbers. This should depress the extent to which participants state blue as the most likely color of a randomly drawn number. When asked to think in terms of font size, we hypothesize that the large words in the decoy distribution will interfere with the recall of (orange) large numbers. This should increase the extent to which participants say small font size is most likely for a randomly drawn number.

We find robust and significant differences in probabilistic assessment between these two cue conditions: the share of participants identifying orange, large numbers as the most likely numbers rises from 17% when asked for size judgments to 40% when asked for color judgments. This evidence supports two important predictions. First, memory cues provided by the question affect recall and probability judgments. Second, biases change as predicted by interference: types occurring frequently in both the target and comparison groups are underestimated. This experiment also suggests that our results are driven by selective recall, not differences in encoding across treatments, as all subjects observe the same set of images under identical conditions.

We see our paper as making two contributions. Relative to the literature on the link between memory and probability judgments, we emphasize the importance of interference in recall. To paraphrase Kahneman, understanding why some thoughts become accessible requires us to understand why others are not. Interference in recall provides a natural mechanism for the key feature of the representativeness heuristic, namely that

beliefs about a group are tilted towards its features that are distinctive *relative* to other groups. In particular, our model predicts when beliefs exaggerate unlikely features, as illustrated in the examples on introverted librarians and oft-divorced Hollywood stars.

Several other studies have linked biases in probabilistic judgments to the properties of memory. Dougherty, Gettys and Ogden (1999) develop the Minerva DM model, building on Hintzman (1984). In this model, the conditional probability $P(t|g)$ is computed by retrieving memory traces that are similar to the cue g, identifying the subset of those that are similar to the target type t, and normalizing. The driving force for biases here is noisy recall, in particular the possibility of false positives driven by partial similarity between cue and items. Types that are more frequent in the broad population may be overestimated for group g because they trigger false positives in this computation. Busemeyer et al. (2011) offer a quantum probability model that also relies on similarity to generate biases. A number of papers develop models of limited sampling from memory which bias the process of searching in memory in various ways, including by anchoring search to an initial cue (Sanborn and Chater 2016) or by focusing on instances associated with salient payoffs (Shi and Griffiths 2009). Crucially, because these models do not allow for interference across groups g, they do not explain why judgments about a target group depend systematically on a comparison group, nor why certain unlikely traits are so often recalled to the detriment of the more common ones.[1]

Other models of memory are motivated 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, 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 activation based models such as the Adaptive Control of Thought - Rational (ACT-R, Anderson and Reder 1999), and inhibition based 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, either through spreading activation or direct inhibition. Instead, the dominant feature of our data is interference, or contrast, *across* groups: if a type becomes more common for other groups, it is less representative of, or less similar to, the target group and it is thus estimated to be less likely.

[1]Dougherty, Gettys and Ogden (1999) propose that base rate neglect reflects an substitution of the question $P(t|g)$ for the questions $P(g|t)$ (see also Frederick and Kahneman 2003). As we discuss in Section 4, this mechanism cannot explain our evidence.

The second contribution is our experimental strategy, which offers some advantages over the classical word-pair paradigm. Using multidimensional objects allows us to create associations in memory, and to study how interference shapes probabilitstic judgments. Specifically, it allows us to naturally represent groups with different distributions of traits. By varying independently the frequency of a type in different groups, we can test how the representativeness of a type in a group depends on its frequency in the other group. In particular, we can recreate the classical patterns of representativeness – overestimation of an unlikely event – with purely abstract groups recalled from memory.

Finally, our results shed new light on the representativeness heuristic. Griffiths and Tenenbaum (2001) offer a formalization of representativeness close to the one we build on, but they do not link it to recall and interference, as we do here. In recent work, we have shown that the statistical notion of representativeness introduced in Gennaioli and Shleifer (GS 2010) and Bordalo, Coffman, Gennaioli and Shleifer (BCGS, 2016) helps account for beliefs in several domains. These range from social stereotypes to the formation of expectations in financial markets settings. In these papers, stereotypes and expectations distort beliefs towards *relatively* more frequent traits of the target distribution. As such, they display a kernel of truth property (Hilton and Von Hippel 1996). The kernel of truth helps account both qualitatively and quantitatively for field data on the belief of individuals about ability across genders (BCGS 2019) and on the expectations of market participants (Bordalo, Gennaioli, Shleifer 2018, Bordalo, Gennaioli, Ma, Shleifer 2018, Bordalo, Gennaioli, La Porta, Shleifer 2018). The results in this paper provide a foundation for the kernel of truth in terms of interference in recall.

Our approach does not cover an important class of phenomena, namely forecasting biases, which are traditionally accounted for by the representativeness heuristic. Examples of such biases include the gambler's and hot hand fallacies and the law of small numbers. As described by Kahneman and Tversky, these biases rely on a notion of representativeness that is different to the one examined here. Specifically, when forecasting the behavior of a random variable, the likelihood of an event "is judged by the degree to which it reflects the salient features of the process by which it is generated" (KT 1972). Because it depends on a mental representation of a data generating process, this notion seems closer to semantic than episodic memory. Understanding the accessibility of thoughts as mediated by semantic memory is an important avenue for future research.

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.

2 Model and Experimental Framework

2.1 The Model

Our model captures the idea that probability judgments are formed using recall of statistical associations, in the sense that a hypothesis is evaluated by selectively retrieving from memory experienced instances of it. Crucially, the retrieval process is subject to interference in ways that we describe below.

The memory database is described by a probability space with event space Ω and probability measure P, which summarizes the entirety of a person's experiences, including their frequencies. Two random variables are defined on this probability space: T, which we think of as types, and G, which we think of as groups. The task of the decision maker is to assess the probability of different types t in T in a certain group g in G.

To give an example, the probability space can contain the distribution of the world population according to a variety of features. The task could be to guess the occupation of an introverted person, in which case the group is $g = introvert$ and subjects assess the probability of different types $T = \{librarian, farmer, \ldots\}$ in g. Alternatively, the group could be a nationality, say $g = Irish$, and the task is to assess the probability of different hair colors, $T = \{dark, light, red\}$, in this group.

A Bayesian retrieves all possible realizations t of T that are consistent with $G = g$ and computes the true conditional probability $P(t|g)$ using the measure P.[2] Relative to this benchmark, in our model specific types t are more easily recalled for group g, and thus their probability is inflated. Formally, we assume that the decision maker assesses the probability of t according to the distorted measure:

$$\tilde{P}(t|g) = P(t|g)\frac{M\left(P(t|g), P(t|-g)\right)}{\mathbb{E}\left(M\right)} \tag{1}$$

The true measure $P(t|g)$ is modified by a factor M which captures the ease with which each type t comes to mind when thinking about group g. Intuitively, Equation (1) says that the easier it is to retrieve t – that is, the higher is M – the more this type is deemed likely. As in Tversky and Kahneman (1972), types that more easily come to mind are judged to be more likely. The ease of retrieval is normalized, so that probabilities add up to one.

Our key assumption is that M increases in the true conditional probability of t in g,

[2] Formally, the agent computes the conditional probability by retrieving all elementary events ω consistent with each realization t and with g and by using the probability measure P to compute: $P(T = t|G = g) = \frac{\int_{\omega\in\Omega:T(\omega)=t,G(\omega)=g} dP(\omega)}{\int_{\omega\in\Omega:G(\omega)=g} dP(\omega)}$.

but decreases in the probability of t in the comparison group $-g$.

$$\frac{\partial M}{\partial P(t|g)} > 0, \qquad \frac{\partial M}{\partial P(t|-g)} < 0,$$

This specification captures two important features of memory. First, it captures the law of frequency, the idea that all else equal more likely types more easily come to mind. This is accounted for in the assumption that M increases in its first argument. When thinking about occupations, $t = teacher$ comes easily to mind because they are so frequent, unconditionally, in the memory database. This relative accessibility implies that, all else equal, judgments overestimate frequent types and neglect infrequent ones.

Second, our model embodies two forms of interference: across groups and across types. Interference across groups means that a type that is common across many groups is less likely to be recalled for each one of them. To continue our example, when cueing the specific group $g = introvert$ the type $t = teacher$ is interfered with because so many instances of teachers are not in the introvert group. This is the key mechanism in our model, formally captured by the assumption that M decreases in its second argument. Often the laws of frequency and interference go against each other, which is precisely when an unlikely type is judged representative. However, they can also reinforce each other, as when a type common in group g is uncommon in other groups.

The other form of interference present in our model is interference across types. This is captured by the normalization factor $\mathbb{E}(M)$, which ensures that probabilities sum to one. The fact that some types are highly accessible, namely they have high M, interferes with recall of other types that are less accessible (Kahana 2012). For instance strong recall of teachers blocks recall of other, less accessible occupations, say librarians.

This memory-based model of probability assessments exhibits three main differences from existing models of recall. First, interference works across groups, not only across types as in the leading models of interference (Anderson and Reder 1999, Dougherty, Gettys, Ogden 1999, Kahana 2012). As we will see, interference across groups is key to understanding why we often overestimate the probability of infrequent outcomes. In fact, the key results of our experiment arise because M decreases in its second argument. Second, in our model encoding is perfect, which is in contrast to some models allowing for noisy retrieval (e.g. 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. If instances of teachers are not stored together with the person's character, the high frequency of the teacher occupation would create strong recall for any group, with little interference from competing groups.

The phenomenon of interference acros s groups can be explained by similarity-based

recall (Kahana 2012). Thinking about the probability of an event (t, g) prompts retrieval of past encounters with identical events (t, g), but of fairly similar events $(t, -g)$ that share the type but not the group. The higher the likelihood of the latter, the more they are likely to be recalled, and thus the greater the interference in the recall of (t, g) itself.[3] In fact, while the model is formally described in terms of probabilistic assessments (Equation 1), the underlying mechanism is one of selective recall from episodic memory. Assessments of probability of a given type should therefore be consistent with assessments of numerosity across types.

To illustrate, take for example the estimation of the frequency of red hair among the Irish. Suppose that $P(red|Irish) = 0.1$ and $P(lightbrown|Irish) = 0.4$ while $P(red|rest\ of\ world) = 0.01$ and $P(lightbrown|rest\ of\ world) = 0.2$, with the others being dark haired. Even though it is rare in absolute terms, red hair is very easy to recall for the Irish because it has virtually no interference from other national groups, while instances of brown haired Irish are strongly interfered with by the much more common brown haired non-Irish. As this example shows, representativeness amplifies true differences between the distributions, a property called the kernel of truth (BCGS 2016, Judd and Park 1993), which may shed light on perceptions about "essential properties" of groups.

The main predictions of our model rely on the general specification in Equation (1) (see Appendix A). In particular, Equation (1) is closely related to KT's definition of the representativeness heuristic, which implies that a type t is more representative for g the higher $P(t|g)$ is *relative* to $P(t|-g)$. That is, when making judgments about statistical associations, what KT call representative types are those that are easy to recall either because they are frequent ($P(t|g)$ is high) or because they face weak interference from a comparison group ($P(t|-g)$ is low) or both. To establish an even tighter connection between memory and representativeness, we can specify that ease of recall is increasing in the likelihood ratio of t in g versus in $-g$:

$$M(P(t|g), P(t|-g)) = \left[\frac{P(t|g)}{P(t|-g)}\right]^{\theta} \tag{2}$$

where $\theta \geq 0$ captures the extent to which ease of recall distorts judgments. This specification of the recall distortion has been used in BCGS (2016) and BGS (2018) as a shortcut for the impact of representativeness on judgments. The model presented here

[3]The formalism of similarity-based recall postulates a similarity relation S between items in memory, so that the likelihood that cue g retrieves type t_i is given by $P(t_i|g) = S(t_i, g)/\sum_j S(t_j, g)$. The normalization factor, which sums over all t_j in memory, captures interference across types, as in the denominator of Equation (1). Interference across groups can instead be captured by the similarity function itself: a type is more similar to a cue g if it is less commonly associated with other groups, as in the function M in Equation (1).

lays out a more general and explicit connection between probability judgments and memory, particularly interference.

2.2 Experimental Design

The Baseline Experiment We now describe our main experiment and the model's predictions. In the next Section, we describe the implementation in detail and present the results.

Participants are shown a sequence of 50 abstract images that vary along two dimensions, content and color. In our baseline experiment, participants 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.

Table 1: Databases of baseline experiment

Image database in the *gray* treatment

		Types:		
		*o*range	*b*lue	*g*ray
Groups:	*n*umbers	10	15	0
	*w*ords	0	0	25

Image database in the *blue* treatment

		Types:		
		*o*range	*b*lue	*g*ray
Groups:	*n*umbers	10	15	0
	*w*ords	0	25	0

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

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 asked:

Q2. How many orange numbers were shown to you?

Q3. How many blue numbers were shown to you?

In both treatments, the target distribution is the distribution of colors of the numbers. Because the numbers presented are the same in both cases, the target distribution is also the same in both cases. But we vary the "decoy" distribution, which consists of either 25 gray shapes or of 25 blue words. This tests the hypothesis that retrieval and probability judgments are swayed, at least in part, by interference. There is more interference along the color dimension when numbers are compared to blue words than to gray shapes. All the experiments presented here build on this basic design, which extends the preliminary experiments in BCGS (2016).

Before moving to the model's predictions, we highlight a few features of our design. First, our design makes use of abstract images to ensure that participants do not have pre-existing associations between features of these images that may distort their assessments, as in Bhatia's (2017) model of semantic associations. Instead, the use of abstract images gives us full control over which associations participants can store and allows us to exogenously vary these associations in the experiment through treatment manipulations, so we can study their causal impact. This enables us to test how participants generate new associations and to explore how these associations are shaped by the statistical properties of the co-occurrence of attributes across groups.

Second, we consider two types of assessments about the target distribution: estimating the likely color and estimating the amount of numbers of each color. In later variations of the experiment, participants also estimate the probability of each color for a randomly drawn number. This provides a within-subject consistency check, and both types of questions are informative about the role of retrieval in the assessment of probabilities.

Third, the flexibility of the experimental design lends itself to extensions and robustness checks. For instance, one might worry that the results are driven by our use of certain types – like colors. We can easily re-craft the abstract images, swapping the colors (see Section 3.3 and, in particular, Figure 3 (a)), or use different image attributes, such as font size (see Section 4 and Appendix C.1).

The Model's Predictions. Participants' memory database consists of 50 images that are characterized by random variables G (content) and T (color). G takes values in $\{N, W, S\}$, for numbers, words, and shapes, while T takes values in $\{o, b, g\}$ for orange, blue and gray. In each treatment, G takes two values $\{g, -g\}$, where $g = N$ and $-g = W$ or S. After having seen the images, participants make a probabilistic assessment about the color distribution of numbers. In the context of the model, they are given the cue $g = N$ and are asked to assess $P(t|N)$. The key outcome of interest is how this assessment depends on the properties of the comparison cue $-g$.

Under the null of perfect retrieval, there would be no treatment effects. A decision

maker with perfect memory asked to assess $P(t|N)$ would retrieve all 50 images from his memory database, compute the relative frequency of orange numbers, and report $P(o|N) = 10/25$. This absence of treatment effects is shared by the much larger class of models of probabilistic judgments in which interference does not play a role. Consider a model of imperfect memory in which decision makers make judgments by sampling from the memory database (Sanborn and Chater 2016). A decision maker who samples randomly sometimes over-samples and sometimes under-samples orange numbers, but on average would produce a correct assessment. Some models assume sampling is guided by payoffs (Lieder, Hsu, Griffiths 2018). While such models may generate predictable biases in probabilistic assessments (e.g., by oversampling states with large payoffs), the biases they generate are driven by the target distribution itself and would be constant across treatments.

If instead retrieval is shaped by interference, then according to Equation (1) participants should overestimate the frequency of colors t that are relatively more likely for numbers than for $-g$. To see this, start with the $gray$ treatment. In this case, orange and blue numbers are both infinitely representative because $\frac{P(o|N)}{P(o|S)} = \frac{10/25}{0/25}$ and $\frac{P(b|N)}{P(b|S)} = \frac{15/25}{0/25}$. Because representativeness distortions are bounded, it follows from Equation (1) that probabilities are not distorted. Intuitively, this is because the target distribution and the decoy distribution do not overlap in the color dimension, so that both colors that occur among the numbers are free from interference and equally likely to come to mind. On average, this yields a correct prediction of $\tilde{P}(o|N) = 10/25$.[4]

Consider now the $blue$ treatment. The key difference relative to the $gray$ treatment is that the decoy distribution now overlaps with the target distribution, because blue words have the same color as blue numbers. With interference in recall, this overlap has a drastic effect. Compared to the $gray$ treatment, orange numbers are still infinitely representative, $\frac{P(o|N)}{P(o|W)} = \frac{10/25}{0/25}$, but the representativeness of blue numbers drops dramatically, $\frac{P(b|N)}{P(b|W)} = \frac{15/25}{25/25}$. Because blue is strongly associated with words, the latter interfere with the recall of blue numbers and reduces their recalled frequency. Equation (1) then yields our main predictions:

Prediction 1. *Assessments of the probability that a random number is orange are greater in the $blue$ treatment than in the $gray$ treatment, formally $\tilde{P}(o|N)_{blue} \geq \tilde{P}(o|N)_{gray}$. In particular, participants are more likely to say that orange is the likely color of a randomly drawn number in the $blue$ treatment.*

To further test the role of interference as captured by the likelihood ratio of the

[4]If in the representativeness expression (2) we take $c > 0$, then the fact that blue numbers are more likely makes them somewhat more representative, which induces-decision makers to inflate their probability. Generally speaking, the BCGS model predicts that in the control group representativeness should distort the assessment in the direction of $\tilde{P}(o|N) < P(o|N)$.

type across the target and decoy distributions, we run additional treatments in which the decoy distribution changes more continuously, by gradually moving from the $blue$ treatment with 25 blue words to treatments that replace some blue words with orange words (for instance, 22 blue words and three orange words). Suppose in treatment $blue_k$ the share of orange words is given by $P(o|W)_{blue_k} = \frac{k}{25}$. Our model predicts that adding orange words increases interference with the recall of orange numbers relative to that of blue numbers. As a consequence:

Prediction 2. *Assessments of the probability that a random number is orange decrease with the number of orange words in the decoy distribution, formally* $\tilde{P}(o|N)_{blue_k} < \tilde{P}(o|N)_{blue_{k'}}$ *for* $k > k'$*. In particular, participants are less likely to say that orange is the likely color of a randomly drawn number as k increases.*

The model's predictions apply to question Q1 about the likely color, which is based on a global assessment of the color distribution of numbers. But as described in Section 2, the model also speaks to how participants estimate the number of images of a given type. Thus, the model predicts that responses to Q1 should be consistent with the responses to Q2 and Q3 (the estimated number of different colors), in the sense that the color with higher estimated number should also be the estimated likely color in Q1. We return to these issues in Section 5.

3 Representativeness and Selective Recall

In this Section we present the results of three sets of experiments: the baseline experiment described in Section 2 (Study 1), a variation in which we add a distraction task before the elicitation of beliefs (Study 1b), and a variation in which we gradually vary the representativeness of colors for numbers (Study 2).

We start by describing in detail the baseline experimental setup. We chose 25 unique numbers between 50 and 70 with precision of one decimal place (for instance, "63.6"). This ensured that each number was distinct, yet likely to be perceived as part of a single group, and not individually memorable. We chose 25 words related to a common theme (time, for instance, "October"), so that no word individually stood out. Finally, we chose 25 shapes from Microsoft PowerPoint (geometric shapes, arrows, etc), whose size we normalized. We used the same numbers, words and shapes in all experiments, and these can be found in Appendix B.1.

At the beginning of the experiment, participants are instructed that they will be shown a sequence of 50 abstract images. They are told they will subsequently be asked

questions about these images and that accurate answers will be compensated.[5] Participants are randomly assigned to treatments. The images are ordered randomly for each participant, each appearing on the screen in isolation for approximately one second. We conjecture that this is enough time for subjects to see each image's t, g type (content and color), but not enough time for them to make detailed notes, take screenshots, or become disengaged. Participants can form a holistic impression of the full distribution they see, as well as memory traces of individual images, but are unlikely to be able to keep a precise tally of each type. After viewing the sequence of images, participants see questions about the images they saw. Appendix B.1 provides details on the materials participants see (including a link to an online version of the blue and gray treatments), as well as on incentives and experimental setup.

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. Study 1 was conducted on MTurk and in the laboratory with $N = 1,013$ in Spring of 2018. Instructions, word images and questions were translated into the appropriate national language. Here we present the aggregated results. We report the disaggregated results as well as procedural details in Appendix B.1.

Predictions and Results The null hypothesis for questions Q1, Q2 and Q3 is that participants have on average the same probabilistic assessment in both the $blue$ and $gray$ treatments. As discussed above, this prediction is shared by several models of limited recall.

In stark contrast, our model (Prediction 1) implies that participants inflate the frequency of orange relative to blue numbers in the $blue$ treatment, $\tilde{P}(o|N)_{blue} > \tilde{P}(o|N)_{gray}$, because recall of blue numbers is interfered with by the presence of blue words in memory. 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, and ii) report a higher share of orange numbers compared to participants in the $gray$ treatment, computed from answers to Q2 and Q3.

The results are summarized in Figure 1 and Table 2 and show strong support for Prediction 1. Column (1) in Table 2 reports an OLS regression of a response dummy (1

[5]Unlike in other memory experiments, such as recall of word pair associations (Kahana 2012), subjects are not told they should memorize the images they see. In lab experiments, participants receive €0.50 per correct answer for those parts of the experiment that are randomly determined to be compensated. In Mturk, most participants receive $0.20 per correct answer. See Appendix B.1 for details.

if "orange is likely") on a treatment dummy (1 if $blue$), 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 20.8pp from the $gray$ treatment to the $blue$ treatment (35.3% to 56%, significant at 1% level).[6]

Column (2) reports the results of an OLS regression of a response dummy that takes value 1 if the participant reported more orange numbers in Q2 than blue numbers in Q3, which is an alternative measure of each participants' belief about the likely color. Consistent with Column (1), there is a 13.8pp increase in the share of participants who recalled more orange than blue numbers in the $blue$ treatment (30.4% to 44.2%, significant at 1% level). Across Columns (1) and (2), the treatment dummy coefficients are close. In fact, answers in Q1, Q2 and Q3 are consistent for roughly 90% of participants. Columns (3) and (4) show how the median quantity of recalled orange and blue numbers depends on the treatment.[7] 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.

Finally, again based on answers to Q2 and Q3, we can compute the ratio of orange numbers to total numbers recalled. Column (5) shows that, as predicted by the model, participants recalled on average a significantly higher share of orange numbers in the $blue$ treatment (50% versus 44.5%). Given that the recalled share of orange numbers is high at baseline, this average increase can have a large effect on the number of participants who say orange is the likely color.

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 (1). The treatment effect represents a jump of over 50% in the frequency of mistakes (Columns 1 and 2), and a 12.5% increase in the estimate of the probability of orange (Column 5), so it is large both in absolute and in relative terms. 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

[6]Nonlinear Logit and Probit regressions yield similar results. In the Appendix we present a variety of robustness checks. The treatment effect is present in all waves and across MTurk as well as the laboratory (see Figure 7 and Tables 5, 6, 7, and 8 in Appendix B.1).

[7]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 Appendix B.1.

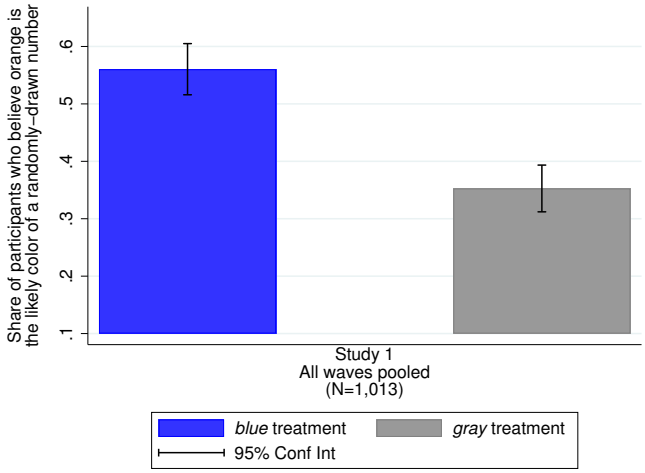


Figure 1: Share of participants who believe that the likely color of a randomly-drawn number is orange for the $blue$ and $gray$ treatments of Study 1.

Table 2: Regression estimates of treatment effects in Study 1

	OLS: Y=1 if "orange is likely"	OLS: Y= 1 if more orange numbers recalled	0.5-Q-Reg Y= Orange numbers recalled	0.5-Q-Reg Y= Blue numbers recalled	0.5-Q-Reg: Y= Share of orange to total numbers recalled
	(1)	(2)	(3)	(4)	(5)
1 if $blue$	.2060***	.1379***	0	− 2***	.0556***
	(.0307)	(.0302)	(.4187)	(.6427)	(.0124)
MTurk dummy	yes	yes	yes	yes	yes
Wave dummies	yes	yes	yes	yes	yes
Constant	.3305***	.3043***	10***	14***	.4444***
	(.0302)	(.0259)	(.3598)	(.5524)	(.0107)
Observations	1,013	1,013	1,013	1,013	1,013
Adj./Ps. R^2	0.04	0.02	0.01	0.01	0.02

with no pre-existing associations in memory.[8,9]

[8]One may ask whether the particular colors chosen impact our results. For instance, a body of work 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.1, that replaced color with font size as the types (i.e. the random variable T), and the results go through (see Table 15).

[9]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. This likely reflects the difficulty of the task and imperfect memory.

3.2 Study 1b: Recall vs. Working Memory

We interpret Study 1 as suggestive of an important role for selective recall in driving biased assessments. But one might ask whether there is a role for working memory in driving our results. As Baddeley (1992) lays out, working memory is the brain system that allows for *temporary* storage and manipulation of information necessary for reasoning. This system is believed to operate separately from the longer-term memory systems that include recall. Could it be the case that the distortions we observe are connected to the working memory systems, rather than recall?

To separate the two, we replicate Study 1 but introduce a distraction task between the viewing of the sequence of images and the answering of the questions about them. The goal of the distraction task is to fully engage participants' working memory in a task orthogonal to the key task at hand (recall of the images). By presenting this distraction task, we likely impose a heavy extraneous cognitive load (see Paas, Renkl, and Sweller 2003), swamping participants' working memories with new information. If our results persist after the distraction, it must be recall rather than working memory.

We used two distraction tasks which varied in content and length. In one version, participants assess the emotional expression in 10 pictures of human faces.[10] Participants needed on average 90 seconds to go through this task. In the other version, participants solve 5 easy raven matrices. Participants needed on average 170 seconds to go through this task. Neither distraction task was incentivized. Study 1b was conducted in the laboratory with $N = 790$ in Spring of 2018. Here we present the aggregated results. We report the disaggregated results as well as procedural details in Appendix B.2.

Predictions and Results. The null hypothesis is that the information participants retrieve during the question stage draws on their long-term memory as opposed to their working memory, and so the introduction of this distraction stage should have no impact on the results of Study 1. Figure 2 shows that Study 1b replicates, both qualitatively and quantitatively, the treatment effect of Study 1 on the share of participants who say "orange is likely".

A regression analysis of the pooled data of Studies 1 and 1b shows that adding the distraction task has no effect on the treatment effect on any of the output variables explored in Table 2 (see Table 10 in Appendix B.2).

Taking stock Studies 1 and 1b sought to maximize the variation in interference across treatments, by minimizing interference in the $gray$ treatment and maximizing it in the

[10]The Emotion Recognition questions are adapted from a quiz created by The Greater Good Science Center at UC Berkeley (https://greatergood.berkeley.edu/quizzes/take_quiz/ei_quiz), as used in Bordalo et al 2018.

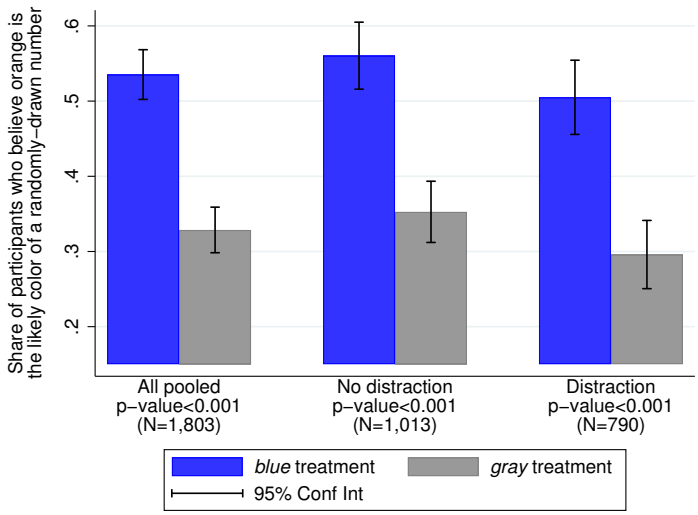


Figure 2: Main treatments w/out and w/ distraction

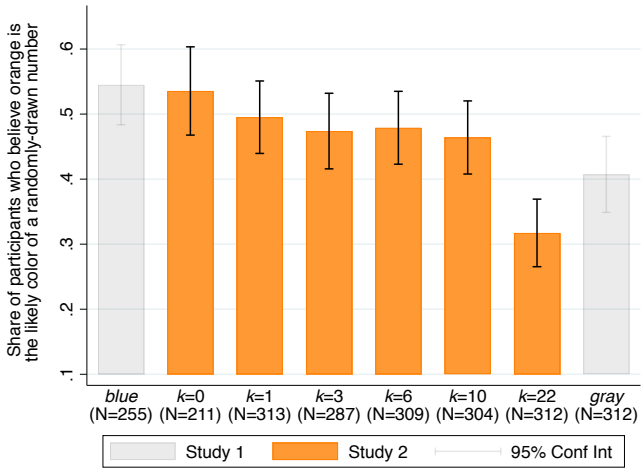
blue treatment. This was reflected both in the choice of decoy content (words may be more similar to numbers than are shapes) and in the choice of color distributions (blue words maximize interference with blue numbers, while gray shapes minimize it). In the next study, we refine our design to explore in more detail the role that relative likelihood plays in shaping distorted recall, as formalized in Equation (2).

3.3 Study 2: Varying Relative Likelihood

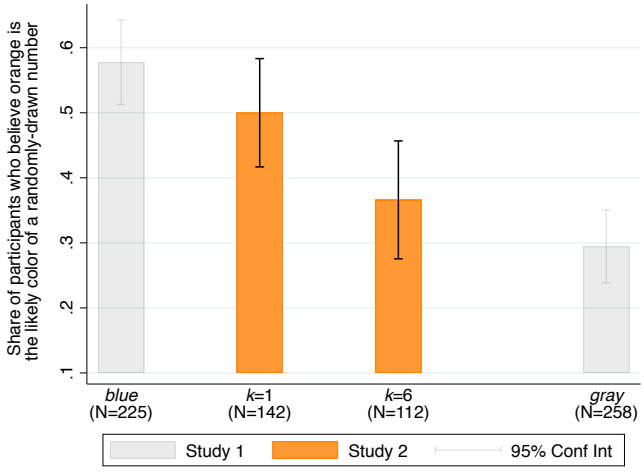
Study 2 differs from Study 1 only in terms of the composition of decoys. As before, the target group is 25 Numbers (10 orange and 15 blue). For the decoy group, we create six variants of the words distribution in the *blue* treatment, denoted $blue_k$, which are characterized by replacing k blue words with orange words, $P(o|W)_{blue_k} = k \in \{0, 1, 3, 6, 10, 22\}$. The six variants range from our original treatment *blue* ($k = 0$) where orange numbers are infinitely representative, to word distributions that are mostly orange, making blue numbers more representative ($k = 22$). Across the variants, the representativeness of orange numbers is decreasing in k. 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.3.

Predictions and Results The null hypothesis is again that participants' recall of the distribution of numbers is on average identical across treatments. Compared to this null, Prediction 2 implies that, as k increases and orange numbers become less representative, fewer participants state orange to be likely and the assessed share of orange numbers decreases. The results are summarized in Figure 3. Here, we separate the MTurk and

laboratory experiments as each explore different values of k.



(a) Conducted with MTurk



(b) Conducted in Laboratory

Figure 3: 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 3 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 $k = 22$.[11]

[11]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

Similar results hold for 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 with the results from Study 1.[12]

Table 3 summarizes the pattern described above. 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 in the amount of orange words they saw. Equivalently, 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 (2) and Column(5), respectively.[13] While the former effect is weakly significant, the latter is significant at the 1% level.

Table 3: Regression estimates of treatment effects in Study 2

	OLS: Y=1 if "orange is likely"	OLS: Y= 1 if more orange numbers recalled	0.5-Q-Reg: Y= Orange numbers recalled	0.5-Q-Reg: Y= Blue numbers recalled	0.5-Q-Reg: Y= Share of orange to total numbers recalled
	(1)	(2)	(3)	(4)	(5)
k (number of orange words)	−.0093*** (.0015)	−.0032** (.0015)	0 (.0468)	.0625* (.0334)	−.0008 (.0006)
Mturk dummy	yes	yes	yes	yes	yes
Constant	.4708*** (.0314)	.3685*** (.0309)	10*** (.9692)	14.625*** (.6913)	.4383*** (.0124)
Observations	2,903	2,903	2,903	2,903	2,903
Adj./Ps. R^2	0.02	0.00	0.00	0.00	0.00

Study 2 provides evidence that the relative frequency of types shapes the magnitude of belief distortions. As we decrease the representativeness of orange numbers by increasing the number of orange words, participants are less likely to recall a randomly-drawn

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 those cases.

[12]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 $k = 10$.

[13]Table 3 shows that, as predicted, the number of recalled blue numbers increases as blue words are replaced with orange words (Column 4). The model similarly predicts that the number of orange numbers should increase. Table 3 shows no effect on the median, but there is a negative effect on the mean, see Appendix B.3.

number as orange, despite the fact that the number of orange numbers is held constant across these variants. The studies above include 8 experiments and a total of 4706 participants, each of whom answered three questions about the distribution of colors of numbers. The picture that emerges is that distortions are supportive of our predictions and that they are large in magnitude, consistently across experiments.

4 Modulating Recall through Cues

In our final experiment, we attempt to better isolate the role for selective recall in driving our results. If probabilistic assessments are based on retrieval, then the cue that triggers retrieval matters. Cues triggering different comparisons should, because of interference, generate different probabilistic assessments.

In the previous experiments, images were characterized by two attributes, G (content) and T (color). The cue $g = numbers$ thus immediately entails a comparison between g and $-g$ along the color dimension. Here we introduce a third dimension T' along which images can vary, which is size. In this setting, a cue consists of a pair (g, t) or (g, t'), which is a group to be assessed along a specific dimension, in this case color or size.

To isolate the effect of cues on assessments, we generate a single set of images such that the two dimensions, color and size, are perfectly correlated within the group of numbers. This ensures that cueing participants about color or size is objectively equivalent, and that in an unbiased benchmark both cues should yield the same assessments. In contrast, changing the cue has significant implications for retrieval if it entails differential interference.

Moreover, 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.[14]

Methods and Predictions This experiment includes numbers and words that vary both in color and in size. This extension adds a new random variable $T' \in \{s, l\}$, where l stands for large and s stands for small. The distribution of colors is, as before, $P(o|N) =$

[14]A precise distinction between selective encoding and selective recall is beyond the scope of this paper. However, it is useful to distinguish a process in which the instability of probabilistic assessments arise 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. The current study supports the former hypothesis.

$\frac{10}{15}$ and $P(b|W) = 1$, but now blue numbers are small, while orange numbers and words are large. The database is:

10 orange, large numbers

15 blue, small numbers

25 blue, large words

Because orange numbers exactly coincide with large numbers, this setup allows us to examine probabilistic assessments about large orange numbers in two different ways. To do so, we run two treatments. In the Color Cue treatment, we exogenously cue participants to think of the images in terms of colors, asking in our main question "What is the likely color of a randomly drawn number?". In terms of the model, "given $G = n$ what is the likelihood of $T = o$" cues participants to recall the color dimension T. Participants then recall the color distribution of numbers, in light of the color distribution of non-numbers. This is shown in the Table below. In this case, the problem becomes

	$T = color$	
	blue	orange
numbers	15	10
words	25	0

similar to the Study 1 task, in which orange numbers are very representative while blue numbers are interfered with by blue words.

In the Size Cue treatment, we instead cue participants to think in terms of the font size of the images, asking as our main question "What is the likely font size of a randomly drawn number?" That is, "given $G = n$ what is the likelihood of $T' = l$", cuing participants to recall the marginal distributions along the size dimension T'. In this case, it is the small numbers that are representative because all words are large and interfere with the recall of large numbers, that is $\frac{P(s|N)}{P(s|W)} = \frac{15/25}{0/25}$ while $\frac{P(l|N)}{P(l|W)} = \frac{10/25}{25/25}$. In this case,

	$T' = size$	
	large	small
numbers	10	15
words	25	0

the framing of the question reverses Prediction 1. We have:

Prediction 3. *Assessment of the probability that a random number is orange/large is higher when cued in terms of color than in terms of size, formally* $\tilde{P}(o|N) > \tilde{P}(l|N)$.

As before, neither the decoy distribution nor the framing of the question should impact the assessments of a decision maker whose recall reflects a fixed, even if biased, representation of the images observed. In this case, probabilistic assessments should not depend on the cue (color or size) adopted. Instead, Prediction 3 holds that overestimation of orange numbers is greater in the color treatment than in the size treatment for the initial two questions.

Methods mirror Study 1, except that the distribution of images is as described above. In addition to the outcome variables that we elicited in Studies 1 and 2, we now include a direct probability measure, as follows:

Q4. [Color Cue] What is the probability that a randomly-drawn number is orange?

Q4. [Size Cue] What is the probability that a randomly-drawn number is large?

These questions allow us to further explore probabilistic judgments and within-subject consistency. We conducted Study 3 in the laboratory with $N = 647$ in Spring and Autumn of 2018. Here we present aggregated results. We report disaggregated results and procedural details in Appendix B.4.

Results Our results provide support for Prediction 3. 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 4 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 4: 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

As in Tables 2 and 3, 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 recalled more orange than blue numbers than more large than small numbers, see Column (2) of Table 4. Second, we look at 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 4.

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 state a significantly higher probability on average 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 4. 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.

These results are inconsistent with an alternative account that in estimating the conditional probability $P(o|N)$ subjects instead report the inverse conditional $P(N|o)$ (Gigerenzer and Hoffrage 1995, Koehler 1996), which would predict a recalled share of orange numbers of 1. This mechanism implicitly assumes that the inverse conditional $P(N|o)$ is more accessible to decision makers than the target one $P(o|N)$. This might be the case when experimental subjects are given $P(N|o)$ and are asked to estimate $P(o|N)$.[15] But it is less likely to be the case here, where participants are first shown the set of numbers and then cued with $g = numbers$ alone. This account is also not consistent with the observation that subjects generate a full probability distribution for numbers in questions Q2 and Q3, nor with the fact that $P(o|N)$ is consistent with the recalled quantity of orange and blue numbers.[16]

[15]The inversion of conditional probabilities may be a valid intuition when subjects are faced with a difficult inference problem and are cued with a quantity – the inverse conditional – that seems a plausible proxy for the result. This intuition may help explain evidence such as the Cab Problem (Kahneman and Tversky 1984), in which subjects say the likelihood that the cab is green given that the witness said it was green equals the unconditional likelihood that the witness is correct.

[16]A related formulation to the inverse conditional that solves the latter problem is a mechanical neglect of base rates, whereby the odds ratio $\frac{P(o|N)}{P(b|N)}$ is assessed as $\frac{P(N|o)}{P(N|b)}\left(\frac{P(o)}{P(b)}\right)^{\gamma}$, with $\gamma < 1$ (Bayes' rule corresponds to $\gamma = 1$). This formulation predicts that if a prior is sufficiently strong, say $P(o|N)$ 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.

5 Discussion

This paper explores the link between intuitive thinking, as described by the representativeness heuristic, and memory. The evidence takes the form of 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 gray shapes. This evidence is consistent with the hypothesis that probabilistic assessments about a group stored in memory involve interference from other groups. The role of interference is most clearly illustrated by the fact that, keeping experience constant, probabilistic judgments about a given group are dramatically altered depending on what comparisons are triggered by the cue (Study 3).

Our experiments help put structure on the statement that the probability of an event "is judged by the degree to which it is similar in essential properties to its parent population" (Kahneman and Tversky 1972). Here, similarity between a trait and a group is captured by the lack of interference with that trait from other groups. This seems consistent with the notion that intuitive judgments of similarity are not merely geometric but rather depend on context. Consider Tversky's (1977) famous finding that Austria is perceived as being more similar to Sweden than to Hungary when Poland is also considered, but is instead perceived as more similar to Hungary than to Sweden when Norway is also considered. Austria shares political traits with Sweden (it is a Western country) and geographic traits with Hungary (it is in Central Europe). The mechanism of interference suggests that when compared to mostly Central European and Socialist countries, Austria's geography is interfered with and its distinctive trait becomes "Western". This raises the perceived similarity with Sweden. In contrast, when Austria is compared with other Nordic countries, its geography is less interfered with (while its Western trait is more interfered with), which pushes similarity judgments to Sweden.

Our experiments suggest that selective retrieval and interference may offer a foundation for the representativeness heuristic, but also more generally for understanding the formation of probability judgments and similarity assessments from experienced statistical associations. In fact, similarity judgments between a type and a cue are captured in our model by the ease-of-recall function M. In our interpretation, both probability and similarity judgments reflect accessibility of information in memory, and are shaped by interference. This approach not only accounts for the close alignment between these two types of judgments observed in the literature (Kahneman and Tversky 1983) but also predicts how such judgments depend on objective properties of the targets being assessed.[17]

[17]In our model, the ranking of types in terms of probability in a group and in terms of representativeness or similarity to a group coincide when the recall distortion is strong, $\theta >> 1$. The two rankings may diverge when the overestimation of representative types is not sufficient to overcome the base rates. For

Another set of important issues concerns the external validity of our experiments. 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. 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, interference across groups 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 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.

Interference can also account for the conjunction fallacy, which is in fact captured by Gennaioli and Shleifer's (2010) model of probability judgments. 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.

6 Conclusion

Our results show that selective retrieval and interference are key mechanisms in the formation of probability judgments from experienced statistical associations. This suggests that memory is a promising setting in which to explore the accessibility of thoughts that drive intuitive thinking (Kahneman 2003). While here we have focused on interference in episodic memory, there are at least two other important features of memory that stand out as being clearly relevant. One is attribute-based similarity: a long literature shows that similarity between cues and items along both intrinsic and contextual dimensions is an important determinant of recall from episodic memory (Kahana 2012). This

example, while subjects say the representative Hollywood actress has divorced several times, they also understand that they are more likely to be Democrat (Kahneman and Tversky 1983).

raises the question of how attribute similarity interacts with interference to shape probabilistic judgments. This question could be addressed in our experimental paradigm by exploiting for instance similarity across colors. Second, in many cases judgments involve a broader representation of the problem. An important class of such problems are forecasting biases, where likelihood of an event "is judged by the degree to which it reflects the salient features of the process by which it is generated" (KT 1972). Pinning down the representation of a problem may be better captured by semantic memory. Extending the framework to incorporate these features may help provide a unified explanation of judgment biases and intuitive thinking.

References

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

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

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

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423.

Benjamin, D. J. (2018). Errors in Probabilistic Reasoning and Judgment Biases. In D. Berheim, S. DellaVigna, & D. Laibson (Eds.), *Handbook of behavioral economics* (Vol. forthcoming, pp. 1–166). Elsevier Press.

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 Economics Review, forthcoming*.

Bordalo, P., Gennaioli, N., LaPorta, R., & Shleifer, A. (2018). Diagnostic expectations and stock returns. *Journal of Finance, forthcoming*.

Bordalo, P., Gennaioli, N., Ma, Y., & Shleifer, A. (2018). *Overreaction in macroeconomic expectations.* Working Paper.

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.* Working Paper.

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.

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

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

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.

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

Hintzman, D. L. (1984). Minerva 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers*, *16*(2), 96–101.

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

Judd, C. M., & Park, B. (1993). Definition and assessment of accuracy in social stereotypes. *Psychological Review*, *100*(1), 109.

Kahana, M. J. (2012). *Foundations of human memory*. OUP USA.

Kahneman, D. (2003). Maps of Bounded Rationality: Psychology for Behavioral Economics. *American Economic Review*, *93*(5), 1449–1475.

Kahneman, D., & Tversky, A. (1972). Subjective Probability: A Judgment of Representativeness. *Cognitive Psychology*, *3*(3), 430–454.

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

Keppel, G. (1968). *Retroactive and proactive inhibition.* TR Dixon & DL Horton (Eds.), Verbal behavior and general behavior theory.

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences*, *19*(1), 1–17.

Lieder, F., Griffiths, T. L., & Hsu, M. (2018). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological Review*, *125*(1), 1.

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

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

Shi, L., & Griffiths, T. L. (2009). Neural implementation of hierarchical bayesian inference by importance sampling. In *Advances in neural information processing systems* (pp. 1669–1677).

Tenenbaum, J. B., & Griffiths, T. (2001). The rational basis of representatives. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 23).

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

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. J. (1957). Interference and forgetting. *Psychological Review*, *64*(1), 49.

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

A Derivation of the Predictions

To map the model to our experimental setting, we focus on the case where there are two types, $T = \{t, -t\}$, and two groups, $G = \{g, -g\}$. Denote the probability distributions as $P(t|G) = P_{t,g}$. Similarly, we use the notation $M_t = M(P_{t,g}, P_{t,-g})$ for the ease-of-recall function M. We first derive the conditions under which judged probability $\tilde{P}_{t,g}$ increases with actual probability $P_{t,g}$, namely:

$$\frac{d\tilde{P}_{t,g}}{dP_{t,g}} = \frac{d}{dP_{t,g}} P_{tg} \frac{M_t}{\mathbb{E}(M)} > 0$$

We have:

$$\tilde{P}'_{t,g} \propto \left(M_{tg} + P_{tg}M'_{tg}\right)\mathbb{E}(M) - P_{tg}M_{tg}\left(M_{tg} - M_{-tg} + P_{tg}M'_{tg} + M'_{-tg}(1 - P_{tg})\right)$$

where we used $\mathbb{E}(M) = P_{tg}(M_t - M_{-t}) + M_{-t}$. The expression above is equal to

$$M_{tg}M_{-tg} + P_{tg}(1 - P_{tg})\left(M'_{tg}M_{-tg} - M_{tg}M'_{-tg}\right)$$

which is positive, because by assumption $M > 0$, $M'_{tg} > 0$ and $M'_{-tg} < 0$.

We next derive conditions under which the judged probability $\tilde{P}_{t,g}$ decreases with interference, that is, with the probability of this type in the comparison group:

$$\frac{d\tilde{P}_{t,g}}{dP_{t,-g}} < 0$$

Using primes to denote derivatives with respect to $P_{t,-g}$, we have

$$\tilde{P}'_{t,g} \propto M'_{tg}\mathbb{E}(M) - M_{tg}\left(P_{tg}M'_{tg} + M'_{-tg}(1 - P_{tg})\right) = (1 - P_{tg})\left(M'_{tg}M_{-tg} - M_{tg}M'_{-tg}\right)$$

which is negative, because by assumption $M > 0$, $M'_{tg} < 0$ and also $M'_{-tg} > 0$ (because M_{-tg} decreases in $P_{-t,-g}$, which in turn decreases in $P_{t,-g}$).

Prediction 1. The frequency of blue objects in the comparison group $-g$ is larger in the $blue$ treatment than in the $gray$ treatment, $P_{b,-g=W} > P_{b,-g=S}$. It then follows from the fact that $\frac{d\tilde{P}_{b,N}}{dP_{b,-g}} < 0$ that $\left(\tilde{P}_{b,N}\right)_{blue} < \left(\tilde{P}_{b,N}\right)_{gray}$ and conversely that $\left(\tilde{P}_{o,N}\right)_{blue} > \left(\tilde{P}_{o,N}\right)_{gray}$.

Prediction 2. As before, $\left(\tilde{P}_{o,N}\right)_{blue}$ decreases with the frequency $P_{o,W}$ of orange words.

Prediction 3. In experiment 3, the type space is two dimensional, $T = \{o, b\} \times \{s, l\}$, with four possible types. However, we hypothesize that the cues $color$ or $size$ restrict

attention to a single dimension of the type space. In this case, the effective state space again has two possible states, and the analysis reduces to the case above.

B Methods, Procedures, and Further Results

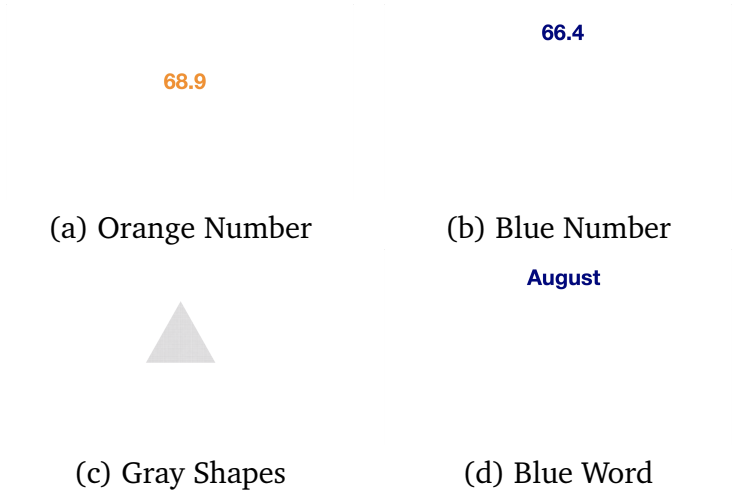
B.1 Study 1

Link to an online version of the two treatments of Study 1:

https://unikoelnwiso.eu.qualtrics.com/jfe/form/SV_6PaEfC4u67hWp1P

Images

Screenshots for each type of image used in Study 1 are displayed in Figure 4. Figures 5 and 6 list the target images (orange and blue numbers) and decoy images (blue words or gray shapes, depending on treatment).



(a) Orange Number (b) Blue Number

(c) Gray Shapes (d) Blue Word

Figure 4: Examples of image screenshots

Questions

- Q1: The computer randomly chose 1 image from all images that were just shown to you. The chosen image showed a number. What is the likely color of the chosen image? Blue or Orange.
- Q2: How many orange numbers were shown to you?
- Q3: How many blue numbers were shown to you?
- *Blue* treatment only:

AddQ1: How many blue words were shown to you?

Orange Numbers:

67.8, **68.9**, **72.6**, **73.0**, **54.0**,
57.1, **65.2**, **52.4**, **56.2**, **61.3**,

Blue Numbers:

66.4, **51.3**, **58.5**, **71.5**, **69.5**,
56.7, **62.7**, **50.8**, **70.8**, **59.6**,
74.3, **53.9**, **63.6**, **60.0**, **64.1**

Figure 5: Target images shown to participants

AddQ2: How many orange numbers were shown to you?

- *Gray* treatment only:

AddQ1: How many gray shapes were shown to you?

Procedural details and data collection

We conducted Study 1 in three waves. Wave 1 was conducted in February of 2018 with MTurk and a sample of 337 participants. Our *blue* and *gray* treatments were accompanied by an unrelated intertemporal choice. The entire experiment lasted for around 10 minutes. Participants received a $1.00 show-up fee. A computer-based coin toss determined randomly whether subjects would receive additional payments based on the *blue* and *gray* treatments or on the unrelated intertemporal choice. In case subjects received additional payments based on the former, one of all participants was randomly chosen to receive $20.00 for each correct answer to Q1-Q3 and AddQ1-2, while all remaining participants received $0.20 for each correct answer.

We then successfully replicated our findings in two more waves, one in the laboratory of the University of Cologne in March of 2018 (N=483) and another with MTurk in March 2018 (N=193). In our first replication, we tested whether our results are robust to moving from MTurk to the laboratory. The laboratory offered us more control on the image display, because we could ensure equally stable internet connection and computing power for each participant. Like in the first wave, the *blue* and *gray* treatments were accompanied by an unrelated intertemporal choice. The entire lab experiment also took 10 minutes. Subjects received a show-up fee of € 4.00. One participant per experimental session (consisting of 26 to 32 participants) was randomly selected to receive additional payments based on the intertemporal choice task. All remaining participants received € 0.50 for each correct answer to Q1-Q3 and AddQ1-2 of the *blue* and *gray* treatments.

Blue Words:

August,	**May**,	**June**,	**July**,	**Millenium**
Hour,	**Month**,	**November**,	**Year**,	**March**
Tuesday,	**December**,	**Second**,	**Saturday**,	**Decade**
January,	**Week**,	**Minute**,	**October**,	**Monday**
Sunday,	**Century**,	**Day**,	**Wednesday**,	**April**

Gray Shapes:

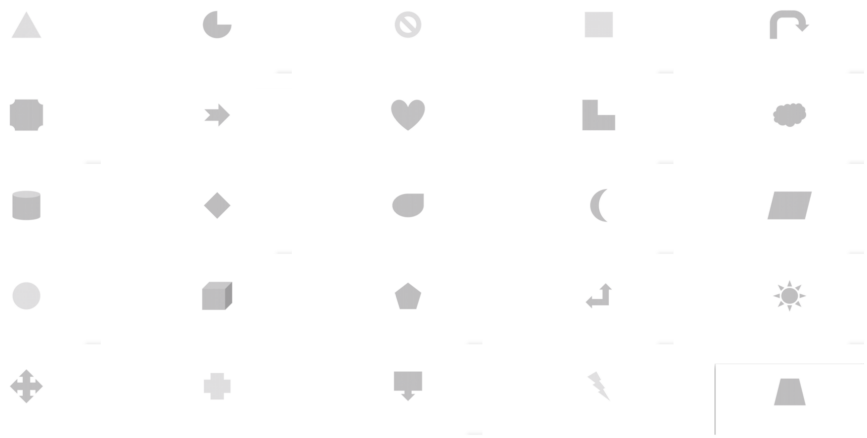


Figure 6: Decoy images shown to participants depending on treatment

In our final replication, we tested whether our results are robust to conducting the *blue* and *gray* treatments without being accompanied by unrelated intertemporal choice tests. The experiment of the third wave took below 7 minutes. Subjects received a $0.50 show-up fee and $0.20 for each correct answer to Q1-Q3 and AddQ1-2.

Further results

Result per wave Figure 1 and Table 2 (see Section 3.1) presented the main findings of Study 1 on the outcomes Q1, Q2, and Q3 which provided support for the predictions of our model.

Figure 7 shows that the share of participants who recalled more orange than blue numbers is larger in the *blue* treatment than in the *gray* treatment for each wave. These differences of 12.71pp, 28.32pp, and 15.03pp for waves 1, 2, and 3, respectively, are significant in OLS regressions (p-values of 0.020, <0.001, and 0.037, respectively). Note that the treatment effect for wave 2—which was conducted in the laboratory—is significantly larger than the pooled treatment effect of waves 1 and 3—which were conducted with MTurk—in a OLS difference-in-differences regression (p-value of 0.018 for the difference-in-differences estimate). Thus, in the laboratory we find a greater treatment

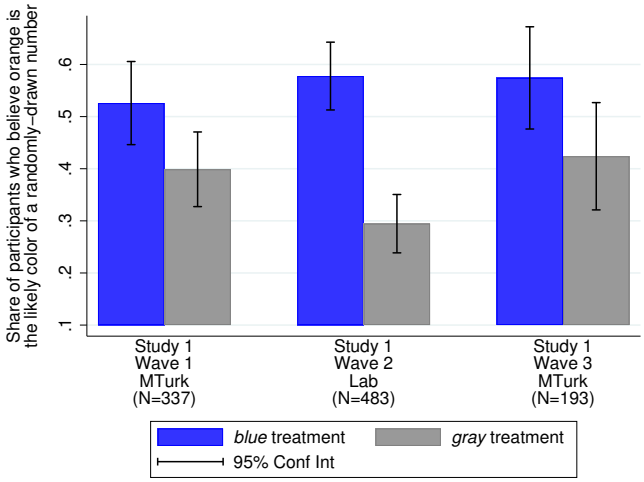


Figure 7: Share of participants who believe that the likely color of a randomly-drawn number is orange for the $blue$ and $gray$ treatments for each wave of Study 1.

effect

Tables 5, 6, and 7 show that the results of Table 2 also hold when focusing only on wave 1, wave 2, and wave 3, respectively. Only the treatment effects on the median orange numbers recalled and median blue numbers recalled do not replicate for wave 3, which has the smallest sample of the three waves.

Table 5: Regression estimates of treatment effects in Study 1 for wave 1

	OLS: Y=1 if "orange is likely"	OLS: Y= 1 if more orange numbers recalled	0.5-Q-Reg Y= Orange numbers recalled	0.5-Q-Reg Y= Blue numbers recalled	0.5-Q-Reg Y= Share of orange to total numbers recalled
	(1)	(2)	(3)	(4)	(5)
1 if $blue$	.1271**	.0758	0	− 2**	.0556***
	(.0542)	(.0527)	(.3916)	(.9791)	(.0100)
Constant	.3989***	.3333***	10***	14***	.4444***
	(.0366)	(.0356)	(.2647)	(.6618)	(.0144)
Observations	337	337	337	337	337
Adj./Ps. R^2	0.02	0.01	0.00	0.01	0.02

Different tests We show in Table 8 that our results presented in Table 2 are robust to using different statistical tests (Logit regressions instead of OLS regressions for outcome measures of Columns 1 and 2 as well as OLS regressions instead of 0.5 quantile regressions for Columns 3, 4, and 5). Participants are significantly more likely to believe

Table 6: Regression estimates of treatment effects in Study 1 for wave 2

	OLS: Y=1 if "orange is likely"	OLS: Y= 1 if more orange numbers recalled	0.5-Q-Reg Y= Orange numbers recalled	0.5-Q-Reg Y= Blue numbers recalled	0.5-Q-Reg Y= Share of orange to total numbers recalled
	(1)	(2)	(3)	(4)	(5)
1 if *blue*	.2832***	.1754	0	− 2***	.0556***
	(.0433)	(.0434)	(.4541)	(.7241)	(.0175)
Constant	.2946***	.2868***	10***	14***	.4444***
	(.0366)	(.0296)	(.3099)	(.4942)	(.0119)
Observations	483	483	483	483	483
Adj./Ps. R^2	0.08	0.01	0.00	0.02	0.02

Table 7: Regression estimates of treatment effects in Study 1 for wave 3

	OLS: Y=1 if "orange is likely"	OLS: Y= 1 if more orange numbers recalled	0.5-Q-Reg Y= Orange numbers recalled	0.5-Q-Reg Y= Blue numbers recalled	0.5-Q-Reg Y= Share of orange to total numbers recalled
	(1)	(2)	(3)	(4)	(5)
1 if *blue*	.1503**	.1521**	2	3**	.0556**
	(.0716)	(.0692)	(1.612)	(1.408)	(.0100)
Constant	.4239***	.2935***	10***	12***	.4444***
	(.0518)	(.0501)	(1.166)	(1.019)	(.0190)
Observations	193	193	193	193	193
Adj./Ps. R^2	0.02	0.02	0.00	0.01	0.03

that a randomly-drawn image is likely to be orange (Column 1). Participants are significantly more likely to recall more orange than blue numbers in the *blue* treatment than in the *gray* treatment (Column 2). Additionally, participants state a significantly greater average share of orange numbers recalled to total amount of images recalled in the *blue* treatment than in the *gray* treatment (Column 5). Columns 3 and 4 show that subjects recall on average more orange numbers and less blue numbers in the the *blue* treatment than in the *gray* treatment, however these differences are not significant.

Table 8: Robustness of regression estimates of treatment effects in Study 1

	Logit: Y=1 if "orange is likely"	Logit: Y= 1 if more orange numbers recalled	OLS Y= Orange numbers recalled	OLS Y= Blue numbers recalled	OLS Y= Share of orange to total numbers recalled
	(1)	(2)	(3)	(4)	(5)
1 if *blue*	.8450***	.5948***	.4667	− .1399	.0341***
	(.0307)	(.0302)	(.4130)	(.5190)	(.0100)
MTurk dummy	yes	yes	yes	yes	yes
Wave dummies	yes	yes	yes	yes	yes
Constant	−.7022***	−.8270***	11.56***	14.06***	.4517***
	(.1137)	(.1162)	(.3549)	(.4461)	(.0144)
Observations	1,013	1,013	1,013	1,013	1,013
Adj./Ps. R^2	0.03	0.02	0.01	0.00	0.01

B.2 Study 1b

Distraction tasks

Emotion expression Figure 8 shows an example of the emotion expression task.

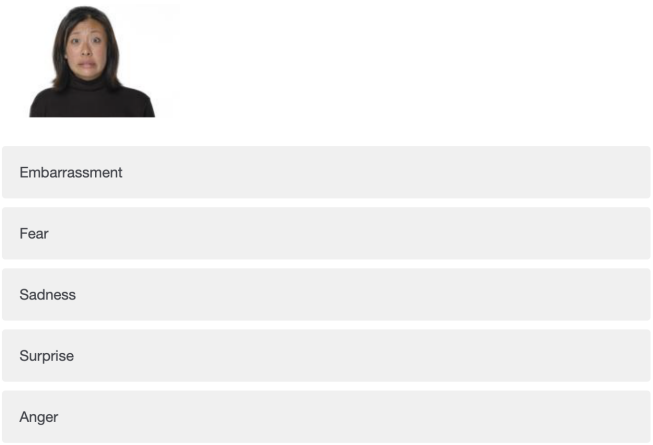


Figure 8: One of the tasks of the emotion recognition questionnaire used to distract participants in wave 1 of Study 1b.

Raven matrices Figure 9 shows an example of the raven matrices task.

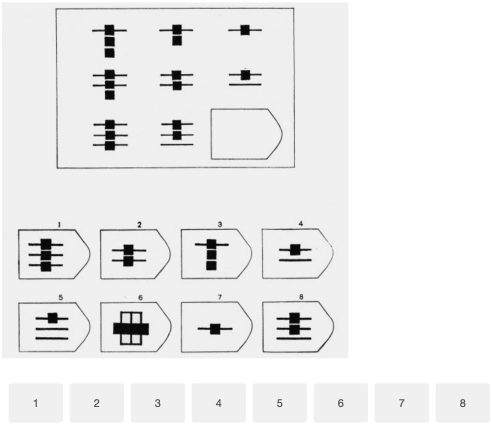


Figure 9: One of the tasks of the raven riddles used to distract participants in wave 2 of Study 1b.

Questions

In order of exposure to participants:

- Q1: The computer randomly chose 1 image from all images that were just shown to you. The chosen image showed a number. What is the likely color of the chosen image? Blue or Orange.
- **Q4**: The computer randomly chose 1 image from all images that were just shown to you. The chosen image showed a number. What is the probability that this number is orange?
- Q2: How many orange numbers were shown to you?
- Q3: How many blue numbers were shown to you?
- *Blue* treatment only:

AddQ1: How many blue words were shown to you?

AddQ2: How many orange numbers were shown to you?

- *Gray* treatment only:

AddQ1: How many gray shapes were shown to you?

- **Q5**: The computer randomly chose 1 image from all images that were just shown to you. The chosen image showed a number. What is the probability that this number is blue?

– The question was used only in the second wave of Study 1b. Because of a computer error, we have responses to this question only for roughly half of the 2nd wave's sample.

Procedural details and data collection

We conducted two waves of the *blue* and *gray* treatments with distraction. They were conducted in May of 2018 in the laboratories of the University of Cologne (N=427) and at Bocconi University (N=363). The memory treatments were accompanied by an unrelated intertemporal choice—like in wave 2 of Study 1 that was also conducted in the lab. The entire lab experiment took 10 minutes. Subjects received a show-up fee of €4.00. In case they were randomly selected to receive additional payments based on our main treatments on memory and representativeness, subjects received €0.50 for each correct answer to the questions on the 50 images. The distraction task on human faces was used in the wave conducted at the University of Cologne and took 90 seconds on average. The distraction task on raven riddles was used in the wave conducted at Bocconi University and took 170 seconds on average. We find no treatment differences between the two waves and hence present the results of the treatments with distraction by pooling both waves. In the following we also show the non-pooled results.

Further results

Results on all outcomes measures of Study 1b: Columns 1, 2, 3, 4, and 5 of Tables 9 show that Study 1b replicates the treatment effects of Study 1 (as presented in Columns 1-5 of Table 2). Additionally, Columns 6 and 7 of Tables 9 show that subjects' direct probabilistic assessment of the likelihood that a random number is orange (Q4) and blue (Q5) differ between the $blue$ and $gray$ treatments as predicted. Participants believe at a random number is more likely to be orange in the $blue$ treatment that in the $gray$ treatment and believe that a random numbers is less likely to be blue in the $blue$ treatment that in the $gray$ treatment.

Differences between *blue* & *gray* w/ and w/out distraction: Tables 10 shows OLS difference-in-difference regressions that test whether the treatment effects of Study 1 and Study 1b differ significantly from each other. The highlighted row shows the difference-in-differences estimates, which are zero or close to zero for all dependent variables in size and do not differ from zero significantly for any of the dependent variables.

Table 9: Regression estimates of treatment effects in Study 1b

	OLS: Y=1 if "orange is more likely"	OLS: Y= 1 if more orange numbers recalled	0.5-Q-Reg Y= Orange numbers recalled	0.5-Q-Reg Y= Blue numbers recalled	0.5-Q-Reg: Y= Share of orange to total numbers recalled	0.5-Q-Reg: Y= Probability that a randomly-drawn number is orange	0.5-Q-Reg: Y= Probability that a randomly-drawn number is blue
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1 if *blue*	.2111***	.1667***	0	− 2***	.0556***	.07***	−.1***
	(.0342)	(.0329)	(.6078)	(.5878)	(.0124)	(.0137)	(.0373)
Wave dummy	yes	yes	yes	yes	yes	yes	no
Constant	.3113***	.2576***	12***	15***	.4444***	.43***	.6**
	(.0284)	(.0273)	(.5046)	(.4880)	(.0114)	(.0144)	(.0285)
Observations	790	790	790	790	790	790	117
Adj./Ps. R^2	0.04	0.03	0.02	0.01	0.02	0.03	0.02

Table 10: Comparing regression estimates of treatment effects between Studies 1 & 1b

	OLS: Y=1 if "orange is more likely"	OLS: Y= 1 if more orange numbers recalled	0.5-Q-Reg Y= Orange numbers recalled	0.5-Q-Reg Y= Blue numbers recalled	0.5-Q-Reg: Y= Share of orange to total numbers recalled
	(1)	(2)	(3)	(4)	(5)
1 if *blue*	.2060***	.1379***	0	− 2***	.0556***
	(.0307)	(.0302)	(.4187)	(.6427)	(.0124)
1 if distraction	− .0549	− .0821**	0	0	−.0159
	(.0409)	(.0398)	(.6253)	(.8114)	(.0164)
1 if *blue* & distraction	**.0051**	**.0288**	**0**	**0**	**0**
	(.0460)	**(.0448)**	**(.7041)**	**(.9134)**	**(.0185)**
MTurk dummy	yes	yes	yes	yes	yes
Wave dummies	yes	yes	yes	yes	yes
Constant	.3305***	.3043***	10***	14***	.4444***
	(.0302)	(.0259)	(.3598)	(.5524)	(.0107)
Observations	1,803	1,803	1,803	1,803	1,803
Adj./Ps. R^2	0.05	0.02	0.01	0.01	0.02

Table 11: Regression estimates of treatment effects in Study 1b for wave 1

	OLS: Y=1 if "orange is more likely"	OLS: Y= 1 if more orange numbers recalled	0.5-Q-Reg Y= Orange numbers recalled	0.5-Q-Reg Y= Blue numbers recalled	0.5-Q-Reg: Y= Share of orange to total numbers recalled	0.5-Q-Reg: Y= Probability that a randomly-drawn number is orange
	(1)	(2)	(3)	(4)	(5)	(6)
1 if *blue*	.2445***	.1615***	0	0	.0455**	.05*
	(.0465)	(.0452)	(.9386)	(.7426)	(.0217)	(.0277)
Constant	.3049***	.2601***	12***	15***	.4545***	.45***
	(.0322)	(.0313)	(.6488)	(.5132)	(.0150)	(.0192)
Observations	427	427	427	427	427	427
(Ps.) R^2	0.05	0.03	0.00	0.00	0.01	0.01

Table 12: Regression estimates of treatment effects in Study 1b for wave 2

	OLS: Y=1 if "orange is more likely"	OLS: Y= 1 if more orange numbers recalled	0.5-Q-Reg Y= Orange numbers recalled	0.5-Q-Reg Y= Blue numbers recalled	0.5-Q-Reg: Y= Share of orange to total numbers recalled	0.5-Q-Reg: Y= Probability that a randomly-drawn number is orange	0.5-Q-Reg: Y= Probability that a randomly-drawn number is blue
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1 if *blue*	.1954***	.1728***	0	− 2**	.0806***	.1***	−.1***
	(.0504)	(.0480)	(.6960)	(.7859)	(.0192)	(.0251)	(.0373)
Constant	.2840***	.2189***	10***	14***	.4194***	.4***	.6***
	(.0368)	(.0351)	(.5088)	(.5745)	(.0141)	(.0184)	(.0285)
Observations	363	363	363	363	363	363	117
(Ps.) R^2	0.04	0.03	0.00	0.01	0.02	0.03	0.02

Differences between the distraction tasks Tables 11 and 12 show treatment effects for both waves in isolation for all outcome measures elicited in that wave, except for the median blue numbers recalled, which only differs between *blue* and *gray* treatments for wave 2 and not for wave 1.

B.3 Study 2

Orange words

In Study 2, participants were exposed to orange words as well as blue words. Some of the words shown in the upper panel of Figure 6 were shown in blue and the remaining were shown in orange.

Procedural details and data collection

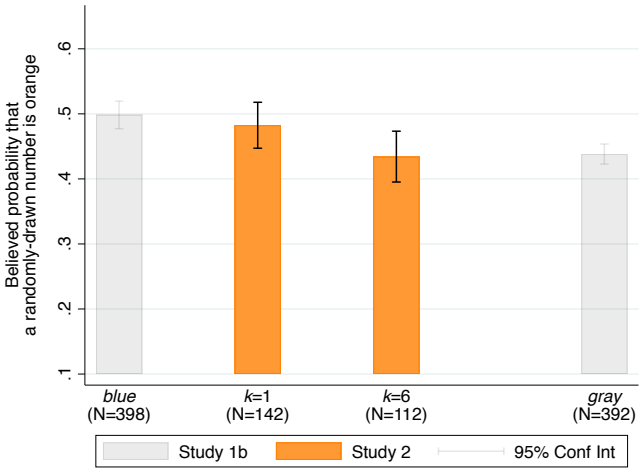


Figure 10: Participants' belief that a random number is orange for the *blue* treatments with $k = 1, 6$.

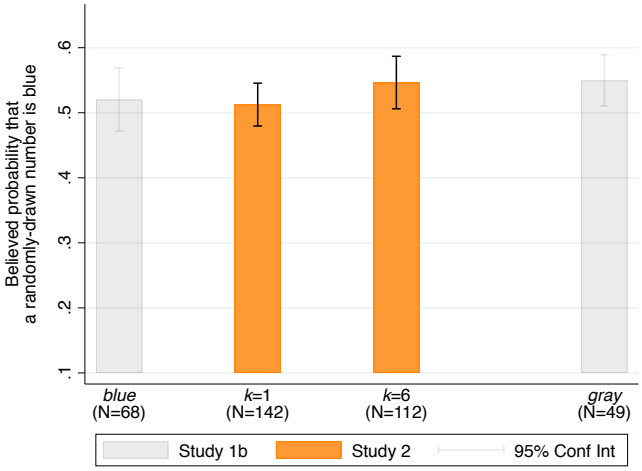


Figure 11: Participants' belief that a random number is blue for the *blue* treatments with $k = 1, 6$.

We conducted three waves of Study 2's treatments. The first two waves were conducted in March and May of 2018 with MTurk, $N = 307$ and $N = 1,431$, respectively.

In both of these waves, the experiment only consisted of the blue treatment variation with orange words. The experiment lasted for 7 minutes for both waves. Participants received a $1.00 show-up fee as well as $0.20 for each correct answer. In the first wave we conducted the treatments with $k = 0, 1, 3, 6$. In the second wave we replicated the first wave and included in addition treatments with $k = 10, 22$.

Table 13: Robustness of regression estimates of treatment effects in Study 2

	Logit: Y=1 if "orange is likely"	Logit: Y= 1 if more orange numbers recalled	OLS: Y= Orange numbers recalled	OLS: Y= Blue numbers recalled	OLS: Y= Share of orange to total numbers recalled
	(1)	(2)	(3)	(4)	(5)
k (number of orange words)	−.0386*** (.0064)	−.0137** (.0064)	−.0288 (.0258)	.0540* (.0289)	−.0023*** (.0006)
Mturk dummy	yes	yes	yes	yes	yes
Constant	− .1141 (.1282)	−.5393*** (.1324)	12.711*** (.5337)	14.945*** (.5985)	.4660*** (.0120)
Observations	2,903	2,903	2,903	2,903	2,903
Adj./Ps. R^2	0.02	0.00	0.00	0.00	0.01

The third wave was conducted in May of 2018 in the laboratory of Bocconi University with $k = 1, 6$. These treatments were accompanied by an unrelated intertemporal choice—like in the laboratory experiments of Studies 1 and 1b. The entire lab experiment took 10 minutes. Subjects received a show-up fee of €4.00. In case they were randomly selected to receive additional payments based on the *blue* treatments with $k = 1, 6$, subjects received €0.50 for each correct answer to the 50 images. We are using the *blue* treatment of the lab experiment of Study 1 and Study 1b as a comparison standard.

Further results

Figures 10 and 11 show that participants' average belief that a random number is orange is greater for fewer orange words and that a random number is blue is lower for fewer orange words. Both findings are consistent with our predictions as discussed in Section 3.3.

We show in Table 13 that our results presented in Table 3 are robust to using different statistical tests (Logit regressions instead of OLS regressions for outcome measures of Columns 1 and 2 as well as OLS regressions instead of 0.5 quantile regressions for Columns 3, 4, and 5). As the number of orange words increases, participants are significantly less likely to believe that a randomly-drawn image is likely to be orange (Column

1); participants are significantly less likely to recall more orange than blue numbers; Additionally, participants state a significantly smaller average share of orange numbers recalled to total amount of images recalled (Column 5). Columns 3 and 4 show that subjects recall on average less orange numbers and more blue numbers as the number of orange words increases, however, only the latter difference is weakly significant.

B.4 Study 3

Questions

Color Cue Treatment (per screen in order of display)

- Screen 1
  - Q1: The computer randomly chose 1 image from all images that were just shown to you. The chosen image showed a number. What is the likely color of the chosen image? Blue or Orange.
- Screen 2
  - Q4: The computer randomly chose 1 image from all images that were just shown to you. The chosen image showed a number. What is the probability that this number is orange?
- Screen 3
  - Q2a: How many blue numbers in small font size were shown to you?
  - Q2b: How many blue numbers in large font size were shown to you?
  - Q3a: How many orange numbers in small font size were shown to you?
  - Q3b: How many orange numbers in large font size were shown to you?

Size Cue Treatment (per screen in order of display)

- Screen 1
  - Q1: The computer randomly chose 1 image from all images that were just shown to you. The chosen image showed a number. What is the likely font size of the chosen image? Small or Large.
- Screen 2
  - Q4: The computer randomly chose 1 image from all images that were just shown to you. The chosen image showed a number. What is the probability that this number is large?
- Screen 3
  - Q2a: How many blue numbers in small font size were shown to you?
  - Q2b: How many blue numbers in large font size were shown to you?
  - Q3a: How many orange numbers in small font size were shown to you?
  - Q3b: How many orange numbers in large font size were shown to you?

Procedural details and data collection

We conducted two waves of Study 3. Wave 1 was conducted in May of 2018 in the laboratory of the Bocconi University ($N = 326$). Wave 2 replicated Wave 1 in October of 2018 in the laboratory of the University of Cologne ($N = 321$). The memory treatments were accompanied by an unrelated intertemporal choice—like in the laboratory experiment of Studies 1 and 1b. The entire lab experiment took 10 minutes. Participants received a show-up fee of € 4.00. In case they were randomly selected to receive additional payments based on the treatments of Study 3, participants received € 0.50 for each correct answer to the questions Q1-Q4.

Further results

We find evidence for our prediction for both Q1 and Q4 in both waves of the Study 3, see Figures 12 and 13.

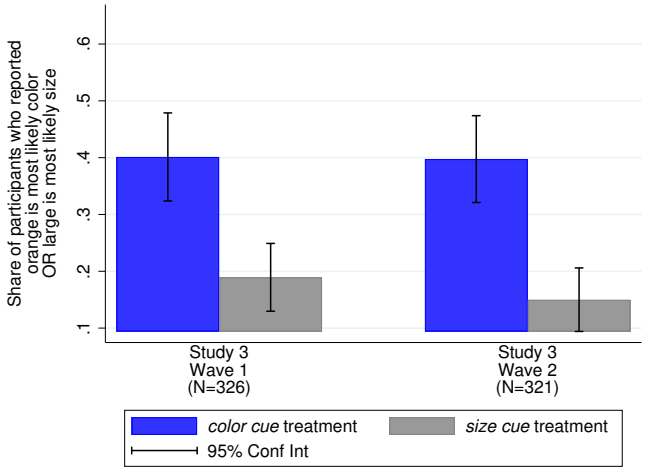


Figure 12: Share of participants who believe that the color OR size of a randomly-drawn number is most likely orange OR large for the treatments of Study 3.

C Further Experiments

In the following, we discuss 4 further experiments that we conducted. The first two, pilot and ProbNumber, provide evidence that the results of Study 1 do not rely on particular design features. In the pilot, for instance, we show that interference-based recall leads to distortions of probabilistic statements when types are constructed around differences in font size (large versus small) rather than color (orange versus blue). The latter two, Studies 2b and 2c, investigates further variations of experimental parameters in order

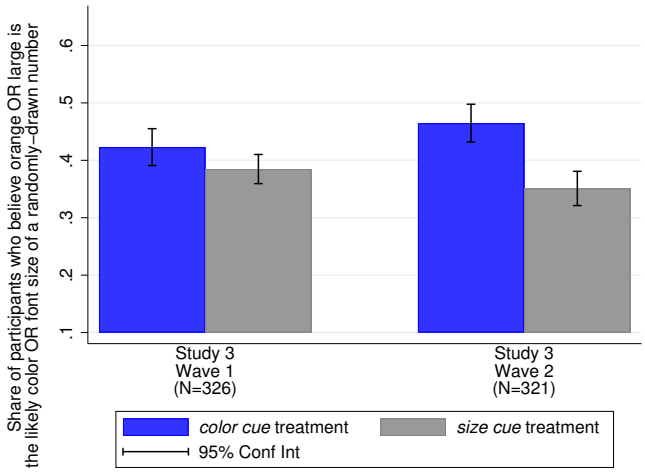


Figure 13: Participants' belief that a random number is orange OR large for the treatments of Study 3.

Table 14: Pilot's *small* treatment and *large* treatment

***Small* (Font Size) Treatment**

	Large	Small
Target Group	10 Large Numbers	15 Small Numbers
Decoy Group		25 Small Words

***Large* (Font Size) Treatment**

	Large	Small
Target Group	10 Large Numbers	15 Small Numbers
Decoy Group	25 Large Numbers	

to seek out the boundaries of our treatments effects and potentially highlight avenues for future research.

C.1 Pilot

The findings of our pilot experiment corroborate the evidence for interference-based distortions of probabilistic judgement from Study 1. Rather than having orange and blue numbers as our target images, we displayed numbers in large and small font size. When large and small numbers are shown to participants along small words, participants think that a random number is more likely to be large than when the numbers appear along large words. Small words seem to interfere with the recall of small numbers, urging

participants to judge that a random number's font size is more likely to be large rather than small. The results of the pilot hence provide further evidence for interference-based probabilistic judgements.

Design

The two between-subjects treatments of our pilot follow the same structure as our baseline treatments of Study 1:

First, participants anticipate to receive questions that are incentivized for accuracy on a sequence of 50 abstract images displayed to them during the experiment.

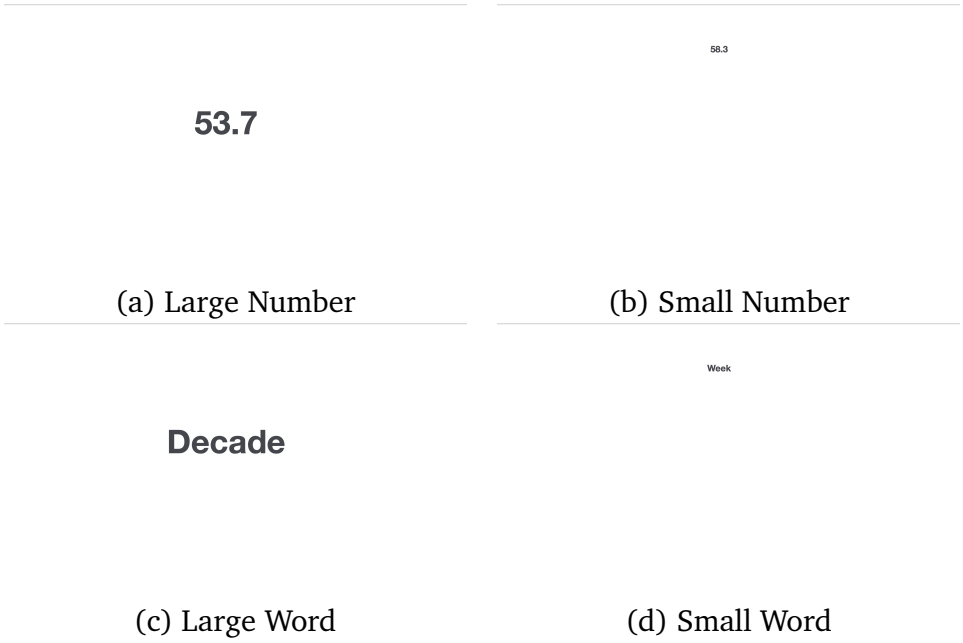
Second, participants see the 50 images that appear on separate screens for short moments of time and in random order. The 50 images vary along two features. The first feature, denoted by G, is the category of the object displayed in the image, which can be a number or word, ie $G \in \{n, w\}$. We will refer to numbers as targets and words as decoys. The second feature, denoted by T, is the font size of the object, which can be large or small, ie $T \in \{l, s\}$. Table 14 shows which types of images participants were exposed for each of the two treatments. Example screenshots for each kind of image are displayed in Figure 14.

Third, participants face questions on the targets which require them to recall the observed sequence of images as well as on the decoys images. The questions—presented here in the same order as show to participants in the experiment—were:

- Q1: The computer randomly chose 1 image from all images that were just shown to you. The chosen image showed a number. What is the likely size of the chosen image? Large or Small.
- Q2: How many large numbers were shown to you?
- Q3: How many small numbers were shown to you?
- Q4: How many small words were shown to you?
- Q5: How many large numbers were shown to you?

Predictions

Our prediction is that as we change the decoy images from being large words to small words, the degree of participants who believe that a randomly-chosen number was shown in large font size should be greater. Because large numbers are representative only in the *small* treatment, interference-based recall inhibits recall of small numbers



(a) Large Number (b) Small Number

(c) Large Word (d) Small Word

Figure 14: Examples of images shown to participants

more in the $small$ treatment than in the the $large$ treatment. Thus, we expect more participants to state that the likely font size is large in the $small$ treatment than in the $large$ treatment.

Procedural details and data collection

We conducted the pilot in December of 2017 with MTurk and a sample of 374 participants. Our *large* and *small* treatments were accompanied by unrelated intertemporal choices. The entire experiment lasted for around 13 minutes. Participants received a $1.00 show-up fee. A computer-based coin toss determined randomly whether subjects would receive additional payments based on the *blue* and *gray* treatments or on the unrelated intertemporal choice. In case subjects received additional payments based on the former, one of every 100 participants was randomly chosen to receive $20.00 for each correct answer to Q1-Q5, while all remaining participants received $1 for each correct answer.

Results

22.4% of participants believe that the likely font size of a randomly-drawn number is large when small and large numbers are shown to participants along large words in the *large* treatment. However, when all words are small in the *small* treatment, 32% of participants believe that the likely font size of a randomly-drawn number is large. This

difference is consistent with our model's prediction and is significantly great than zero in an OLS regression, see Table 15 Column 1.

Mirroring our results from Study 1, we also find that a greater share of participants recalls more large than small numbers in the *small* treatment than in the *large* treatment, Column 2 of Table 15. Additionally, participants recalled a weakly significantly greater median share of large numbers to the total amount of numbers in the *small* treatment than in the *large* treatment, Column 3 of Table 15. However, we also find that the median amount of recalled large numbers as well as the medial of recalled smaller numbers do not differ across treatments.

Table 15: Regression estimates of treatment effects in the pilot

	OLS: Y=1 if "orange is likely"	OLS: Y= 1 if more orange numbers recalled	0.5-Q-Reg Y= Orange numbers recalled	0.5-Q-Reg Y= Blue numbers recalled	0.5-Q-Reg Y= Share of orange to total numbers recalled
	(1)	(2)	(3)	(4)	(5)
1 if *blue*	.1006**	.0883**	0	0	.0274*
	(.0460)	(.0402)	(.8609)	(.9589)	(.0157)
Constant	.2240***	.1421***	8***	15***	.3571***
	(.0329)	(.0287)	(.6152)	(.6852)	(.0112)
Observations	374	374	374	374	373
Adj./Ps. R^2	0.01	0.01	0.00	0.00	0.00

C.2 Study ProbNumber

Design

The two between-subjects treatments of our Study ProbNumber follow the same structure as our baseline treatments of Study 1:

First, participants anticipate to receive questions that are incentivized for accuracy on a sequence of 50 abstract images displayed to them during the experiment.

Second, participants see the 50 images that appear on separate screens for short moments of time and in random order. The 50 images vary along two features. The first feature, denoted by G, is the category of the object displayed in the image, which can be a number, word or shape, ie $G \in \{n, w, s\}$. The second feature, denoted by T, is the color of the object, which can be blue or orange, ie $T \in \{b, s\}$. Table 16 shows which types of images participants were exposed for each of the two treatments. Example screenshots

Table 16: Study ProbNumber's *numbers* treatment and *shapes* treatment

***Numbers* Treatment**

	Decoy Group	Target Group
Numbers	10 Orange Numbers	15 Blue Numbers
Non-numbers		25 Blue Words

Shapes Treatment

	Decoy Group	Large Group
Numbers		15 Blue Numbers
Non-numbers	10 Orange Shapes	25 Blue Words

for each kind of image are displayed in Figure 15. In the ProbNumber study, blue objects are the target images and the orange objects are the decoys.

Third, participants face one question on the targets which require them to recall the observed sequence of images:

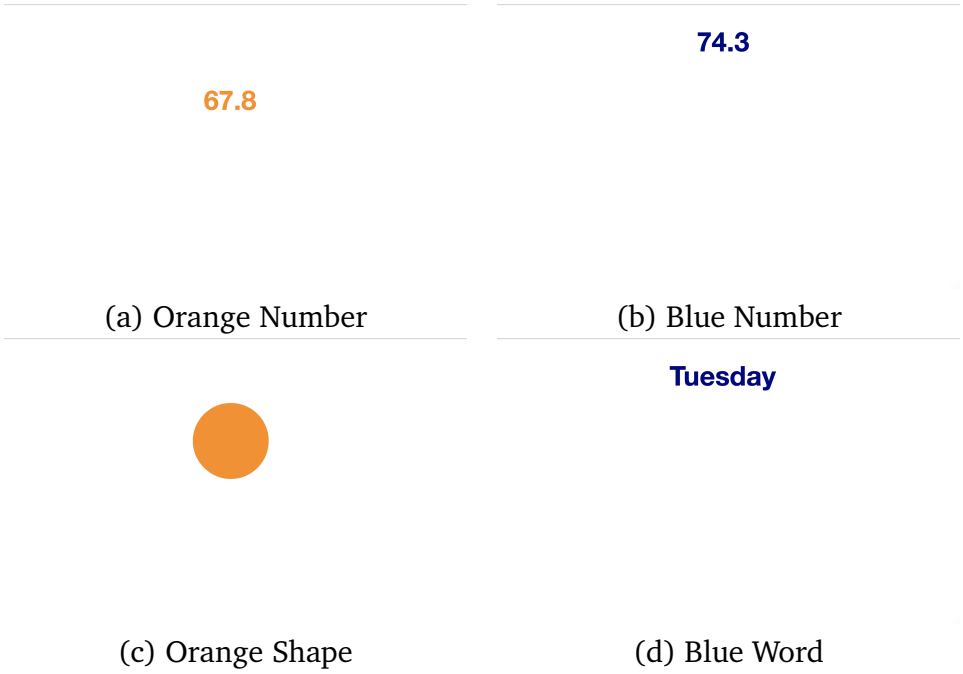
- Q1: The computer randomly chose 1 image from all images that were just shown to you. The chosen image showed a blue object. What is the probability that the chosen image is a number?

Prediction

Our prediction is that as we change the decoy images from being orange shapes to orange numbers, the likelihood that participants recall a randomly-chosen blue object as begin a number should be lower. Because orange numbers are representative only in the *numbers* treatment, interference-based recall inhibits recall of blue numbers more in the *numbers* treatment than in the the *shapes* treatment. Thus, we expect participants to state lower probabilities that a randomly-chosen blue object is a number in the *numbers* treatment than in the *shapes* treatment.

Procedural details and data collection

We conducted the ProbNumber Study in March of 2018 with MTurk and a sample of 304 participants. The experiment consisted only of the *numbers* and *shapes* treatments. The entire experiment lasted for around 7 minutes. Participants received a $1.00 show-up fee and $1.00 for a correct answer to Q1. Because of a computer error, we had to



(a) Orange Number (b) Blue Number

(c) Orange Shape (d) Blue Word

Figure 15: Examples of images shown to participants in Study ProbNumbers

drop one observation. Including an observation with any type of response to Q1 does not alter our results.

Results

Consistent with our prediction, we find that participants believe that a randomly-drawn blue object is a number to a greater extend in the *shapes* treatment—where orange shapes are unlikely to interfere with the recall of blue numbers—than in the *numbers* treatment—where orange numbers are predicted to interfere with the recall of blue numbers.

The median probability that a blue object is a number is 50% in the *shapes* treatment and 40% in the *numbers* treatment. This difference is significant at the 1% level in a 0.5 quantile regression. We find a similar treatment effect when looking at the mean instead of the median. The mean probability that a blue object is a number is 48% in the *shapes* treatment and 42% in the *numbers* treatment. This difference is significant at the 5% level in an OLS regression.

The results of ProbNumber provide further evidence for how interference-based recall drives distortions of probabilistic judgements.

C.3 Study 2b

In Study 2b, we build on Study 1 and Study 2 by further varying the composition of images in order to explore the role of relative likelihoods in driving distorted retrieval more directly.

Design

In Study 2b, we vary likelihood ratios by changing the composition of images within our target group, the Numbers. Consider first our baseline paradigm of Study 1, which consists of two treatments: the *blue* treatment group, which pairs the Numbers with the decoy group of 25 Blue Words, and the *gray* control group, which pairs the same Numbers group with instead a decoy group of 25 Gray Shapes. Study 2b parallels that design, pairing a target group with one of two decoy groups: either 25 Blue Words or 25 Gray Shapes. But, in Study 2b, we vary the target group of Numbers, essentially creating a 3 x 2 design. Table 17 presents the full 3 x 2 treatments. In the first case, we use a target group of 12 orange numbers and 13 blue numbers, pairing it either with the 25 Blue Words (to create $blue\ i = 13$) or the 25 Gray Shapes (to create $gray\ i = 13$). The second case uses a target group of 5 orange numbers and 20 blue numbers, again paired with either the blue words ($blue\ i = 20$) or the gray shapes ($gray\ i = 20$). Finally, in the last set of treatments, we use 2 orange numbers and 23 blue numbers, paired with either blue words ($blue\ i = 25$) or gray shapes ($gray\ i = 25$).

Questions in order of display

- Q1: The computer randomly chose 1 image from all images that were just shown to you. The chosen image showed a number. What is the likely color of the chosen image? Blue or Orange.
- Q2: How many orange numbers were shown to you?
- Q3: How many blue numbers were shown to you?
- *Blue* treatments only:

AddQ1: How many blue words were shown to you?

AddQ2: How many orange numbers were shown to you?

- *Gray* treatments only:

AddQ1: How many gray shapes were shown to you?

Table 17: Study 2b's *Blue i* Treatment and *Gray i* Treatment

***Blue i* Treatments**
(with i $\in \{13, 20, 25\}$)

	Orange	Blue	Gray
Target Group	25—i Orange Numbers	i Blue Numbers	
Decoy Group		25 Blue Words	

***Gray i* Treatments**
(with i $\in \{13, 20, 25\}$)

	Orange	Blue	Gray
Target Group	25—i Orange Numbers	i Blue Numbers	
Decoy Group			25 Gray Shapes

Predictions

Our prediction is that as we increase the share of blue numbers, the likelihood that participants recall a randomly-chosen number as orange should decrease in the *blue* treatment. Thus, we expect a smaller treatment effect (difference between *blue* and *gray*) as the share of blue numbers increases. Denote the number of blue numbers as i, so that the number of orange numbers is 25-i and assume $c > 0$. Then, representativeness-based recall yields the following prediction:

As the share of blue to orange numbers increases, the share of participants who believe that the likely color of a random number is orange should decrease in the *blue* treatment, because the assessed probability that a random number is blue increases, formally $\tilde{P}(b|n)_{blue\ i} \geq \tilde{P}(b|n)_{blue\ i'}$ for $i > i'$.

Procedural details and data collection

We conducted two waves of Study 2b. The first wave was conducted in March of 2018 via MTurk with 601 participants. The experiment consisted only of our memory treatments. The experiment lasted for 7 minutes. Participants received a $1.00 show-up fee as well as $0.20 for each correct answer. We then replicated these results in the laboratory of the University of Cologne with 516 participants. The memory treatments were accompanied by an unrelated intertemporal choice—like in the laboratory experiment of Study 1. The entire lab experiment took 10 minutes. Participants received a show-up fee of € 4.00. In case they were randomly selected to receive additional payments based on our treatments on memory and representativeness, participants received € 0.50 for

each correct answer to the questions on the 50 images.

Results

Our findings provide evidence for our prediction. When the share of blue to orange numbers is increased, participants' *blue* to *gray* treatment difference in their assessed probability that a random number is orange decreases. The treatment effect on the share of participants believing that "orange is more likely" is 15.3pp for 13 blue numbers and 12 orange numbers. The treatment effect reduces to 10pp for 20 blue and 5 orange numbers and to 6pp for 23 blue and 2 orange numbers. While all three treatment effects are (at least weakly) significantly different from zero, the former one is larger than the latter two. OLS regressions show that this difference in treatment effects is at most weakly significantly different from zero when comparing the 13 blue numbers and 12 orange numbers case with the 23 blue numbers and 2 orange numbers case as well as when comparing the 13 blue numbers and 12 orange numbers case with the pooled cases of 23 blue numbers and 2 orange numbers as well as 20 blue numbers and 5 orange numbers. Column (1) of Table 18 shows the results of an OLS regression of the latter result. The weakly significant interaction term (Row (3)) implies the discussed difference in treatment effects. Column (2) of Table 18 shows that the difference in treatment effects increases in size and significance when the main treatments of Study 1 are included: The treatment effect for treatments with 20 or 23 blue numbers is significantly smaller than for treatments with 13 or 15 blue numbers.

Studies 2b provide further evidence that likelihood ratios are directly linked to the extent of distortion in recall. As we increase the representativeness of blue numbers in the *blue* treatment, the size of the treatment effect when comparing across blue words and gray shapes is directionally decreased. Thus, it seems clear that likelihood ratios have a large role to play in predicting the accuracy of recall.

C.4 Study 2c

In Study 2c, we build on Study 1 by further varying the composition of images in order to explore the boundaries of how many decoy images are needed to bring about interferference-based recall. While this test lays outside of our mode, we take it as an instructive direction for future research on improved models of interferference-based recall.

Table 18: Regression estimates of treatment effects in Study 2b

	Including only Study 2b OLS: Y=1 if "Orange is more likely"	Including Study 1 & 2b OLS: Y=1 if "Orange is more likely"
	(1)	(2)
1 if *blue*	.1669***	.1957***
	(.0411)	(.0237)
1 if i (Blue Numbers) $\geq$20	−.3044***	−.2897***
	(.0354)	(.0397)
1 if *blue* & $i \geq 20$	−.0904*	−.1192***
	(.0499)	(.0656)
MTurk dummy	yes	yes
wave dummies	–	yes
Constant	.3302***	.3153***
	(.0325)	(.0300)
Observations	1,117	2,130
Adj. R^2	0.18	0.15

Notes: For Study 2b, wave dummy and MTurk dummy are collinear

Design

In Study 2c, we vary the amount of decoy images. Consider first our baseline paradigm of Study 1, which consists of two treatments: the *blue* treatment group, which pairs the Numbers with the decoy group of 25 Blue Words, and the *gray* control group, which pairs the same Numbers group with instead a decoy group of 25 Gray Shapes. Study 2c parallels that design, pairing a target group with one of two decoy groups: either Blue Words or Gray Shapes. But, in Study 2c, we vary the amount of decoys, essentially creating a 4 x 2 design. Table 19 presents the full 4 x 2 treatments. In the first case, we show the target group of 10 orange numbers and 25 blue numbers, pairing it either with the 5 Blue Words (to create $blue\ j = 5$) or the 25 Gray Shapes (to create $gray\ j = 5$). The second case uses the same target group of 10 orange numbers and 15 blue numbers, paired with either 50 Blue Words ($blue\ j = 50$) or 50 Gray Shapes ($gray\ j = 50$). In the latter two cases, the target group is paired with 75 Blue Words ($blue\ j = 75$) or 75 Gray Shapes ($gray\ j = 75$) or alternatively with 125 Blue Words ($blue\ j = 125$) or 125 Gray Shapes ($gray\ j = 125$).

Questions in order of display

Table 19: Study 2b's *Blue j* Treatment and *Gray j* Treatment

***Blue j* Treatments**
(with j $\in \{5, 50, 75, 125\}$)

	Orange	Blue	Gray
Target Group	10 Orange Numbers	15 Blue Numbers	
Decoy Group		j Blue Words	

***Gray j* Treatments**
(with j $\in \{5, 50, 75, 125\}$)

	Orange	Blue	Gray
Target Group	10 Orange Numbers	15 Blue Numbers	
Decoy Group			j Gray Shapes

- Q1: The computer randomly chose 1 image from all images that were just shown to you. The chosen image showed a number. What is the likely color of the chosen image? Blue or Orange.
- Q2: How many orange numbers were shown to you?
- Q3: How many blue numbers were shown to you?
- *Blue* treatments only:

AddQ1: How many blue words were shown to you?

AddQ2: How many orange numbers were shown to you?

- *Gray* treatments only:

AddQ1: How many gray shapes were shown to you?

Procedural details and data collection

We conducted two waves of Study 2c. The first wave was conducted in March of 2018 with MTurk and a sample of 800 participants featuring all treatment cells of the 4 × 2 design. We then replicated the *blue* treatments with 5, 50, and 75 decoys in May of 2018 with MTurk and a sample of 592 participants. In both waves, the experiment consisted only of our memory treatments. The experiment lasted for 7 minutes. Participants received a $1.00 show-up fee as well as $0.20 for each correct answer to the questions on the 50 images.

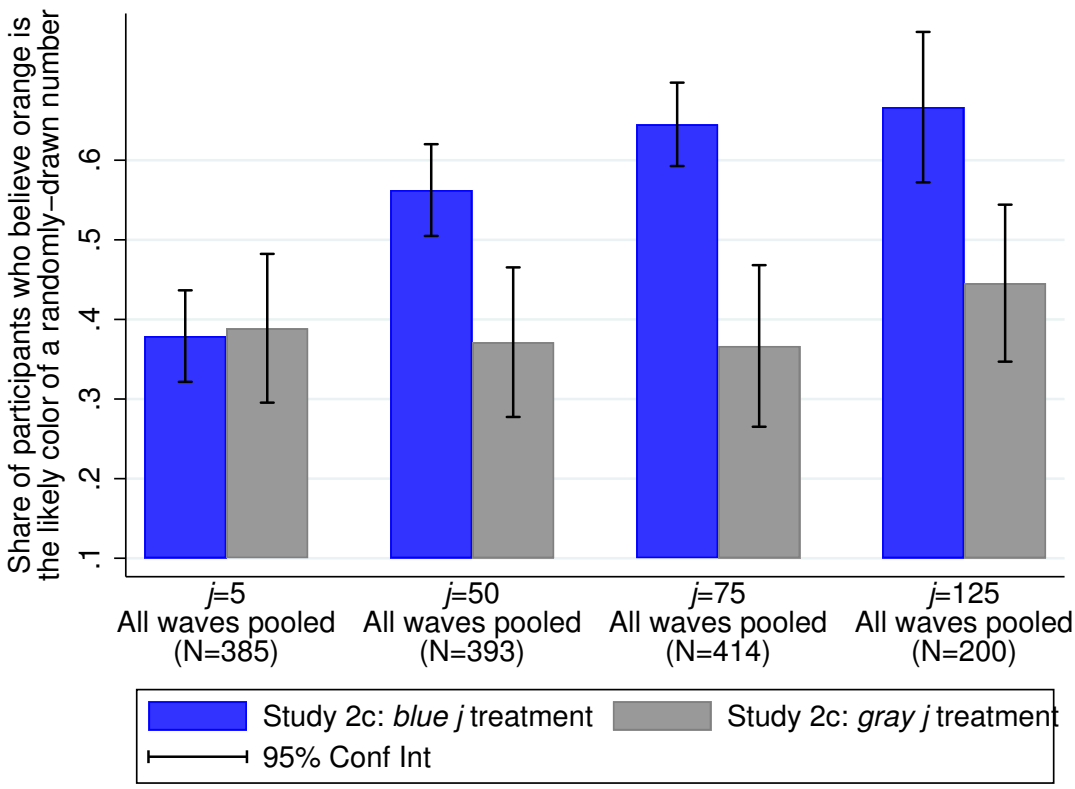


Figure 16: Vary the amount of decoy images

Results

Figure 16 shows that we find no treatment effect when the blue and orange numbers are displayed along 5 decoys (in the *blue* and the *gray* treatments). For 50, 75, and 125 decoys, however, we do find significant treatment effects that resemble our findings of Study 1. These findings suggest a lower bound for decoys to be able to interfere with target images.