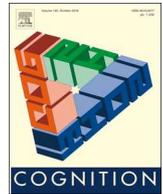




ELSEVIER

Contents lists available at ScienceDirect

Cognition

journal homepage: [www.elsevier.com/locate/cognit](http://www.elsevier.com/locate/cognit)

COGNITION

Original research article

## A dilution effect without dilution: When missing evidence, not non-diagnostic evidence, is judged inaccurately

Adam N. Sanborn<sup>a,\*</sup>, Takao Noguchi<sup>b</sup>, James Tripp<sup>a</sup>, Neil Stewart<sup>a</sup><sup>a</sup> University of Warwick, UK<sup>b</sup> University College London, UK

When people make important decisions, such as determining the guilt or innocence of a defendant, it is desirable to have multiple sources of information. For example, jurors in criminal cases and auditors determining the likelihood of fraud both make critical decisions and will want to have as much information about their case as possible. However, much research has shown that using more than one source of evidence results in biased judgments of the combined evidence, especially when one of the sources of evidence is non-diagnostic. With non-diagnostic evidence people show a *dilution effect*: despite its irrelevance, inclusion of non-diagnostic evidence tends to weaken or dilute diagnostic evidence. For example, in an experiment one group of mock jurors were asked if a man had murdered his aunt, and were told the diagnostic information that he was known to have argued with his aunt and had no alibi. A second group were asked the same question, and were told the same diagnostic information, but were additionally told the nondiagnostic information that he was of average height and had average vision. This second group were less confident that the man was guilty of murder: the non-diagnostic information had diluted the diagnostic information [Zukier and Jennings \(1984\)](#).

The dilution effect has been found in many different tasks, including tasks with objective probabilities that ask participants to reason about the likelihood that a set of poker chips were drawn from one of two book bags [Labella and Koehler \(2004\)](#), [Shanteau \(1975\)](#), [Troutman and Shanteau \(1977\)](#), tasks that ask for predictions about the behavior of people [Nisbett et al. \(1981\)](#), [Zukier \(1982\)](#), tasks asking for consumers' judgments of products [Meyvis and Janiszewski \(2002\)](#), and even perceptual tasks [Hotaling et al. \(2015\)](#), [Yurovsky et al. \(2013\)](#). High-stakes judgments are not immune to the dilution effect. The dilution effect is exacerbated by participants being held accountable [Tetlock and Boettger \(1989\)](#), [Tetlock et al. \(1996\)](#), the dilution effect has been found in hypothetical judgments of guilt by mock jurors [Zukier and Jennings \(1984\)](#), and professional auditors show the dilution effect when judging the likelihood of fraud both in the laboratory [Hackenbrack \(1992\)](#) and in their professional judgments [Waller and Zimelman \(2003\)](#).

The dilution effect is also not an isolated cognitive bias – it has similarities to other known judgment biases. One well-known and

related bias is the conjunction fallacy, in which participants judge the probability of one statement to be less than the probability of the conjunction of this same statement with another statement [Tversky and Kahneman \(1983\)](#). Following the example above, a conjunction fallacy would be judging the probability that a man had murdered his aunt and was of average height to be greater than the probability that a man had murdered his aunt – as the first statement is more restrictive, according to the laws of probability it cannot also be more probable. The conjunction fallacy task can require participants to judge situations that are very similar to those in our dilution effect example, but it differs in the type of judgment participants are asked to make.

Another related bias is the latent scope bias, which occurs when people judge a “narrow-scope” explanation, that is one with fewer unverified predictions, as more likely than a “broad-scope” explanation, even in experiments in which the diagnostic evidence is missing [Khemlani et al. \(2011\)](#). As an example of the latent scope bias, you might be asked whether a man had murdered his aunt or not, and told that if he had murdered his aunt there would be shell casings on the floor and his customary mug of hot chocolate would have been left on the table (broad scope), but if the second suspect had done it there would only be shell casings (narrow scope). In this question the shell casings are non-diagnostic as they do not discriminate between the suspects, and the mug of hot chocolate is diagnostic because they do. Despite not knowing whether the diagnostic mug of hot chocolate is present, people tend to think the second suspect (narrow scope) is more likely. This bias is related to the dilution effect in that it compares explanations with diagnostic and non-diagnostic evidence, though both the setup and the question asked of participants is different.

These related cognitive biases share many of the same explanations, explanations which aspire to be general explanations of how people make judgments. One explanation that has been advanced to explain the dilution effect, conjunction fallacy, and latent scope bias is that people make their judgments based on representativeness rather than probabilities. Representativeness assumes that people judge the strength of the evidence by calculating the similarity between the judgment to make and the evidence presented. This has often been cast as a feature matching process in which the features that are in common

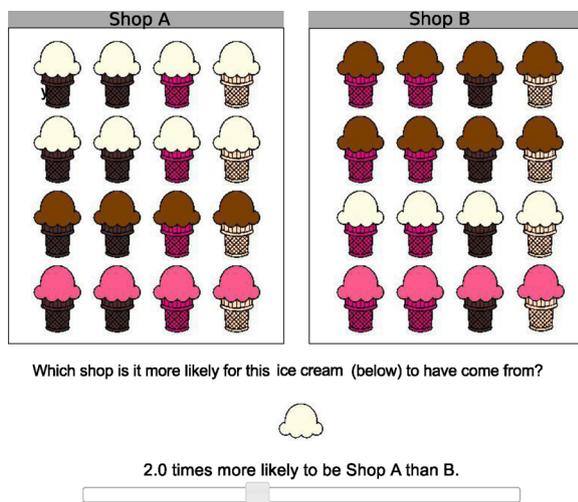
\* Send Correspondence To: Department of Psychology, University of Warwick, Coventry CV4 7AL, UK.

E-mail address: [a.n.sanborn@warwick.ac.uk](mailto:a.n.sanborn@warwick.ac.uk) (A.N. Sanborn).

<https://doi.org/10.1016/j.cognition.2019.104110>

Received 4 October 2018; Received in revised form 11 October 2019; Accepted 14 October 2019

0010-0277/ © 2019 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).



Trial Type	Example Item	Correct Response
Diagnostic		Is 2 times more likely to be Shop A than B
Diagnostic + Non-Diagnostic		Is 2 times more likely to be Shop A than B

Dilution Effect is shown when:



Fig. 1. Example of the task (left) and key test stimuli (right) in Experiment 1a.

between the target and the judgment to make count in favor, while the features that are distinctive between the target and the judgment to make count against. For the dilution effect juror example above, the diagnostic information that the man was known to argue with his aunt and had no alibi increases the similarity of the man to the judgment of guilt, while the non-diagnostic information that the man is of average height and has average vision are distinctive from the features commonly associated with a murderer, and thus reduce the similarity between the man and the judgment of guilt. Therefore, representativeness predicts a dilution effect [Kahneman and Tversky \(1972\)](#), [Nisbett et al. \(1981\)](#), [Zukier \(1982\)](#).

A second explanation, proposed for both the dilution effect and conjunction fallacy, is that people use an averaging rule instead of the correct combination rule when making a judgment. For the dilution effect juror example, averaging the strength of the diagnostic evidence (e.g., arguing with his aunt and having no alibi) with the strength of the non-diagnostic evidence (e.g., average height and vision) produces a lower average strength compared to the diagnostic evidence alone, and thus produces a dilution effect [Anderson \(1967, 1981\)](#).

Other explanations have been advanced to explain just the dilution effect. One of these explanations is that participants are performing a biased hypothesis test in which they count confirming evidence in favor of a focal hypothesis and both non-diagnostic and disconfirming evidence against the focal hypothesis [Meyvis and Janiszewski \(2002\)](#). Applied to our example, the diagnostic evidence of the man arguing with his aunt and having no alibi counts in favor of the focal hypothesis of guilt, but the non-diagnostic evidence of being average height and having average vision counts against the hypothesis of guilt, and so produces a dilution effect. While experiments have shown that the dilution effect is complex and very likely has multiple causes, all of the above hypothesized causes of the dilution effectively pin the blame on the combination of non-diagnostic evidence with diagnostic evidence, and indeed the name “dilution effect” does so as well. This leads to an interesting question for the dilution effect, which in past work focused on the differences between judgments of diagnostic evidence alone (D) and judgments of diagnostic plus non-diagnostic evidence (D + ND). Is the error in the D + ND judgments, leading to an underestimation of the evidence relative to a normative standard as the name “dilution effect” suggests? Or is it possible that the dilution effect is instead caused by a bias in judging the D evidence, leading to an overestimate of this evidence rather than underestimate of the D + ND evidence?

Here we use a paradigm with objective evidence that allows us to separately determine the accuracy of D and D + ND judgments. Using objective evidence allows us to compare judgments against an objectively correct answer, and thus to assess the overall accuracy of the

judgment for D evidence and the judgment for D + ND evidence, as well as the difference between them. Most previous research on the dilution effect has used subjective rather than objective evidence, which means whilst the difference between D and D + ND evidence can be assessed, the overall levels of each cannot. Examples have included using personality characteristics as evidence when predicting a student’s future grade point average [Zukier \(1982\)](#), or using a face morph as evidence when judging which of the two real faces the morph resembles more [Hotelling et al. \(2015\)](#).

The few studies that used objective evidence all used classic “book bag and poker chip” tasks. In book bag and poker chip tasks, participants are informed of the proportion of poker chips of various colors in each of two book bags and then are asked to judge the probability that a sequence of poker chips came from one bag instead of the other. However, the empirical evidence for the dilution effect in this task is both mixed [Troutman and Shanteau \(1977\)](#), [Labella and Koehler \(2004\)](#) and potentially confounded: participants might not weigh the evidence presented early in a sequence the same as evidence presented later in a sequence, so any dilution effect found could instead be due to primacy or recency effects [Wallsten \(1976\)](#).

To avoid the difficulties of subjective evidence and sequential effects, we designed an objective task that presented evidence simultaneously, by breaking the two pieces of evidence into two separate components: an ice cream flavor and a cone flavor. In our task, participants were asked to judge whether a particular ice cream, particular cone, or particular ice cream and cone was purchased from one of two ice cream shops. To objectively assess the evidence provided by the ice cream, cone, or ice cream cone, participants were shown what the two shops had sold during that day (see [Figure 1](#)). The content of the shops determined whether ice creams and cones were diagnostic or non-diagnostic. By presenting the ice cream and cone simultaneously, we avoid any confounding because of sequential effects.

Below, we report a series of experiments using these stimuli. In Experiments 1a, 1b, 2, and 3 we find that D + ND judgments are closer to normative than D judgments. This surprising result is not anticipated by any of the existing explanations of the dilution effect, or indeed by the name “dilution effect” itself. Our main result was robust: it held for different ways of presenting the background information, evidence, and for both likelihood and probability response scales. We next introduce a new explanation for the dilution effect, which is a modification of a recently proposed explanation for the latent-scope bias: that participants notice that there is missing information in the D stimuli and preferentially “fill in” this missing information in a biased fashion, which generally results in an overestimate of the strength of the D evidence. This new explanation of the dilution effect predicts a

manipulation for decreasing the size of the effect in certain situations, which we experimentally verify in Experiment 4. Finally, we explore how our findings can explain previous empirical puzzles involving the dilution effect, discuss why multiple explanations are needed for the dilution effect, consider how filling-in could make sense as a computational strategy, and conclude with the wider implications for participants filling in missing information.

## 1. Experiments 1a and 1b: Combined ice cream cones in shops

We begin with two experiments (Experiment 1a and 1b) in which participants were asked to respond on a likelihood ratio scale, which can reduce the effect of response biases like conservatism compared with using a probability scale Phillips and Edwards (1966). The correct likelihood ratio was easy to calculate: it was the number of matching stimuli in one shop divided by the number of matching stimuli in the other shop. That is, there is a simple strategy for producing normative responses using the likelihood scale because the correct answer can be found just by counting the ice creams cones in each shop that match the particular ice cream, cone, or ice cream and cone target. Experiment 1a used a 2:1 evidence strength for diagnostic evidence, and Experiment 1b used a 3:1 evidence strength for diagnostic evidence.

### 1.1. Methods

#### 1.1.1. Participants

Participants were recruited from Amazon Mechanical Turk, and were recruited at different times for the two evidence strength experiments. We recruited 114 participants for Experiment 1a. The 104 (43 female, 60 male, 1 unreported) participants who finished this experiment had a mean age of 34.7 (SD = 10.6). We recruited 105 participants for Experiment 1b. The 97 (40 female, 56 male, 1 unreported) participants who finished this experiment had a mean age of 32.8 (SD = 9.0). Participants in each experiment received \$1 as payment on completion of the experiment.

#### 1.1.2. Materials

Participants were shown a pair of ice cream shops, labelled Shop A and Shop B. Each shop consisted of a set of complete ice cream cones: 16 per shop in Experiment 1a and 25 per shop in Experiment 1b. In both experiments, ice cream cones were composed of one of three flavors of ice cream (vanilla, strawberry, and chocolate) and one of three flavors of cone (plain, strawberry, and chocolate) though these flavors were never referred to by name. One of the three flavors of ice cream and one of the three flavors of cone provided diagnostic evidence for Shop A, another flavor of each provided diagnostic evidence for Shop B, and the third provided non-diagnostic evidence. In Experiment 1a, the count of the three flavors of ice creams and three flavors of cones were 8:4:4 in Shop A and 4:8:4 in Shop B. In Experiment 1b, these counts were 15:5:5 in Shop A and 5:15:5 in Shop B. The ice cream and cone flavors were combined factorially such that the evidence provided by each ice cream or cone flavor was independent of the other component. Thus, in Experiment 1a each diagnostic ice cream or cone flavor provided 2:1 evidence, while in Experiment 1b each diagnostic ice cream or cone flavor provided 3:1 evidence, whether they were alone or combined with a non-diagnostic piece of evidence. If both ice cream and cone flavors were diagnostic toward the same shop, then the combined evidence was 4:1 in Experiment 1a and 9:1 in Experiment 1b. If both ice cream and cone were diagnostic toward different shops, the combined evidence was always 1:1. Flavors were randomly assigned to these roles for each participant, with independent randomization for ice cream and cone flavors. For each individual participant, the shop names and their contents were constant throughout the experiment.

Participants made responses on a slider scale that ranged in equal steps from 10 times more likely for Shop A to 10 times more likely for Shop B. The current position of the slider was labelled “X times more

likely to be Shop A than B.” Responses could also go one final step further in either direction which was labelled, “More than 10 times more likely to be Shop A than B” or vice versa. The slider began each trial in the middle of the scale labelled, “Equally likely to be Shop A and B”. Once participants were satisfied with the position of the slider, they pressed a button to submit their response. The scale was discretized to one decimal point.

A mistake in the experimental code allowed participants to select values such as “Shop A is 0.5 times more likely than Shop B” between “Equally likely to be Shop A and B” and “Shop A is 1 times as more likely than Shop B”, which is easy to misinterpret. We corrected for this error by coding all responses between “Shop B is 1 times more likely than Shop A” and “Shop A is 1 times as more likely than Shop B” as responses stating that the shops were equally likely. This affected 3.2% of responses across Experiments 1a and 1b. All of the analyses below on the corrected responses were also run with the uncorrected responses, and the overall conclusions were the same.

#### 1.1.3. Procedure

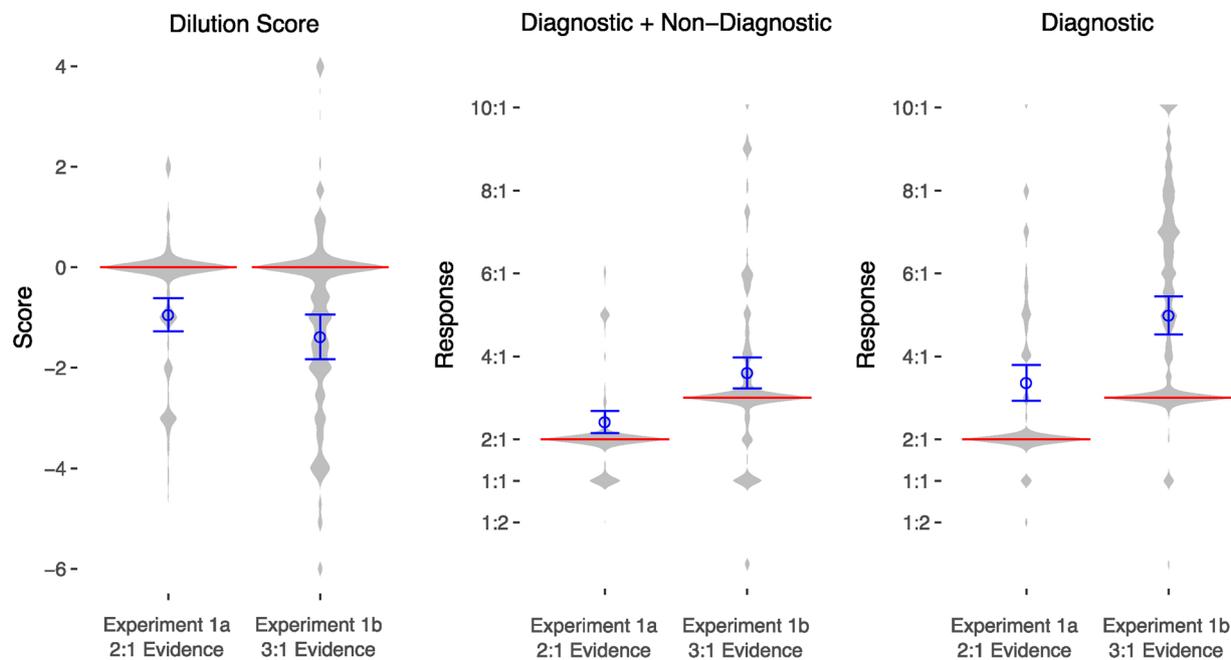
In both experiments, participants were instructed that there were two ice cream shops in town, Shop A and Shop B, and that each shop sold three kinds of ice cream and three kinds of cones. The shop displays were described as a summary of what the owners of each ice cream shop had sold that day. During the task participants were asked to imagine that they were walking around town and had to figure out whether a person had been to Shop A or Shop B based on what ice cream, cone or ice cream cone they were holding. To prepare participants for the different types of stimuli, they were told that sometimes they could see both the ice cream and the cone, and other times they could see only the ice cream or only the cone. They were told to make their judgement in terms of how much more likely it was that this person went to one shop or the other.

After receiving instructions, participants were asked to judge the likelihood ratio of 15 stimuli: each of the 3 cones and 3 ice creams alone, and each of the 9 possible combinations of ice creams with cones. Stimuli were presented in a new random order for each participant.

### 1.2. Results

Because the results of the two experiments were very similar, we interleave their results here. To determine whether we had found a dilution effect in our novel task, we first calculated a dilution score by subtracting each D response from each D + ND response whenever the D stimulus was the same between the two (e.g., subtracting the response to a vanilla ice cream from the response to a vanilla ice cream on a plain cone in Figure 1). This score was a difference in likelihood ratio responses, where a negative value indicated the dilution effect. Because the scale was discretized, it was possible for participants to produce a dilution score that was exactly correct, despite the potential imprecision introduced by participants using a pointing device. The plot of these scores in the left panel of Figure 2 shows that the correct score was the most likely score, 43% in Experiment 1a and 25% in Experiment 1b, but that errors tended in the direction of a dilution effect: the D responses were more extreme than the D + ND responses. In Experiment 1a 73% of errors were in the direction of the dilution effect, and in Experiment 1b this percentage was 75%.

Throughout this paper, we assess the evidence using both null hypothesis significance testing (e.g., t-tests) and Bayesian hypothesis tests. For the Bayesian hypothesis tests,  $BF_{10}$  refers to the Jeffreys-Zellner-Siow (JZS) Bayes factor in favor of the alternative hypothesis over the null hypothesis using the BayesFactor R package default of  $r = 0.707$  for t-tests Rouder et al. (2009). A value of  $BF_{10}$  above 1 is evidence for the alternative hypothesis, while a value below 1 is evidence for the null hypothesis. The value of  $BF_{10}$  itself gives the evidence, but for ease of interpretation values of 3, 10, and 100 have been respectively termed substantial, strong, and decisive evidence for the alternative hypothesis,



**Fig. 2.** Results of Experiments 1a and 1b. The means in the Dilution Score plot are equal to the means in the Diagnostic + Non-diagnostic plot minus the means in the Diagnostic plot. Negative values of the dilution score indicate the dilution effect. The red (dark grey) horizontal lines show the normative response in each plot. The blue (dark grey) circles indicate the means across participants, with the error bars showing the 95% confidence intervals of the means. The light grey regions within each plot give Gaussian kernel density estimates (with bandwidth of 0.1) of the of the raw dilution scores and trial-by-trial responses pooled across participants, with the width of each region normalized to its maximum value.

while values of 1/3, 1/10, and 1/100 have been respectively termed substantial, strong, and decisive evidence for the null hypothesis (e.g., Wetzels et al. (2011)).

In Experiment 1a, we found evidence for a dilution effect as the mean dilution score was below zero,  $M = -0.94$ ,  $SD = 1.69$ ,  $t(103) = -5.71$ ,  $p < .001$ ,  $BF_{10} = 113127$ . The dilution effect was also found in Experiment 1b,  $M = -1.38$ ,  $SD = 2.20$ ,  $t(96) = -6.20$ ,  $p < .001$ ,  $BF_{10} = 789353$ .

We also compared the responses to D stimuli to the objective correct answer and responses to the D+ND stimuli to the objective correct answer. Following the hypothesis implicit in the name “dilution effect”, we would expect the D+ND responses to be on average underestimated and the D responses on average to be accurate. We found a different result. In Figure 2, the D responses were overestimated to a greater extent than the D+ND responses in both experiments. In Experiment 1a the mean D response was higher than the normative value  $M = 3.36$ ,  $SD = 2.23$ ,  $t(103) = 6.23$ ,  $p < .001$ ,  $BF_{10} = 1.1 \times 10^6$ , and the mean D+ND response was also higher than the normative value  $M = 2.42$ ,  $SD = 1.39$ ,  $t(103) = 3.06$ ,  $p = .003$ ,  $BF_{10} = 8.63$ . In Experiment 1b the D response mean was higher than the normative value  $M = 4.99$ ,  $SD = 2.27$ ,  $t(103) = 8.64$ ,  $p < .001$ ,  $BF_{10} = 6.0 \times 10^{10}$ , and the D+ND response mean was also higher than the normative value  $M = 3.60$ ,  $SD = 1.87$ ,  $t(96) = 3.19$ ,  $p = .002$ ,  $BF_{10} = 12.4$ .

### 1.3. Discussion

Overall, we found a very reliable dilution effect in both experiments for both levels of evidence, but we did not find that the D responses were accurate while the D+ND responses were diluted. Instead we found that the D responses were clearly overestimated, and the D+ND responses were also overestimated but to a lesser extent.

These overestimates cannot be explained as the usual conservative bias of participants using a likelihood response scale (e.g., Phillips and Edwards (1966)). If D+ND trials were actually judged accurately, a conservative response bias would have brought the responses below the normative standard of 2:1, while we found that they were above the

normative standard.

However, there is a potential confound in Experiments 1a and 1b. The stock in each of the shops, as shown in Figure 1, were presented as ice cream flavors combined with cone flavors, which matched the way that the stimuli were presented when participants made D+ND judgments. In contrast, the D evidence stimuli were presented differently than the stimuli in the shops: only an ice cream alone or a cone alone were presented. It could be that the D judgments were less accurate because the task of evaluating the evidence was harder; unlike the D+ND stimuli, the D stimuli forced participants to filter out the irrelevant aspect of the ice cream cone.

## 2. Experiment 2: Separate ice creams and cones in shops

In Experiment 2, we created a new display for the ice cream cones in the Shops in order to control for the potential confound in Experiments 1a and 1b. Instead of presenting the ice cream and cone flavors combined into ice cream cones, we separated out the ice creams sold during the day from the cones sold during the day (see example of Shops C and D in Figure 3). This display reverses the relative difficulty of evaluating the D and D+ND evidence in Experiments 1a and 1b. In this experiment it is easier to assess the D evidence than the D+ND evidence because the D stimuli match how the stimuli were displayed in the shops, while evaluating the D+ND evidence requires combining evidence together.

### 2.1. Methods

#### 2.1.1. Participants

We recruited 113 participants from Amazon Mechanical Turk, and 102 completed the experiment (51 female, 49 male, and 2 unreported). The average age was 34.4 ( $SD = 11.1$ ). Participants were paid \$1 upon completion.

#### 2.1.2. Materials

As in Experiment 1a, participants were shown a pair of ice cream

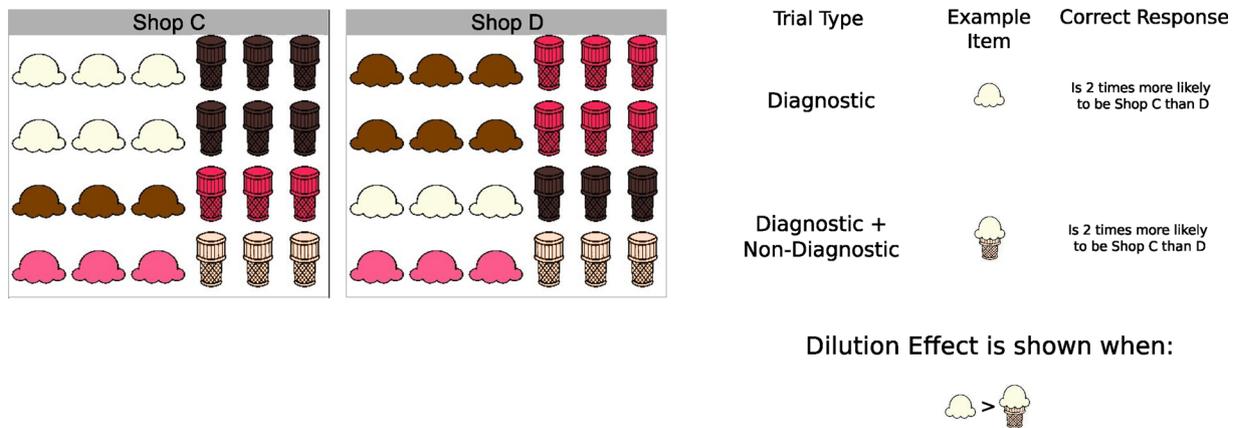


Fig. 3. Example of pairs of shops shown to participants in Experiments 2 and 3 in which the ice creams were separated from the cones.

shops, labelled by letters, containing three flavors of ice cream and three flavors of cone in a ratio of 6:3:3 in one shop and 3:6:3 in the other shop. There were 12 examples of ice cream and 12 examples of cones in each shop, arranged separately instead of combined, as shown in Shops C and D of Figure 3. The types and amount of evidence and the response scale used were the same as in Experiment 1a. We corrected for the programming error in the scale in the same way it was corrected for in Experiments 1a and 1b, and this affected 7.3% of responses.

2.1.3. Procedure

Participants received the same instructions as in Experiments 1a and 1b, with the additional instruction to emphasize the independence of how the ice creams and cones were combined, “The owners of the shops did not tell you which ice creams were sold with which cones. However they did say that people choose ice creams and cones independently: their choice of ice cream will tell you nothing about their choice of cone.” As in Experiments 1a and 1b, participants judged all of the 15 possible stimuli.

2.2. Results

As in Experiments 1a and 1b, we calculated a dilution score for each pair of D and D + ND responses, by subtracting the D response from the D + ND response. This score was a difference in likelihood ratio responses, where a negative value indicated the dilution effect. Figure 4 shows the distribution of these individual dilution scores. While the correct score was most likely, representing 27% of all scores, 66% of errors showed the dilution effect. Using the mean score for each participant, we found evidence for a dilution effect as the dilution score was reliably below zero  $M = -0.44, SD = 0.98, t(101) = -4.51, p < .001, BF_{10} = 919$ .

The D + ND stimuli appeared to be more difficult in this experiment than in Experiments 1a and 1b, but we found the same qualitative result: D + ND responses were overestimated less than D responses. The overall mean response was closer to normative for the D + ND responses, as can be seen in Figure 4. Both responses were on average above the normative value: the mean D response was higher than the normative value  $M = 3.31, SD = 1.96, t(101) = 6.77, p < .001, BF_{10} = 1.1 \times 10^7$ , as was the mean D + ND response  $M = 2.87, SD = 1.89, t(101) = 4.66, p < .001, BF_{10} = 1639$ .

2.3. Discussion

In this experiment, we again found that participants showed a dilution effect, but that the D + ND evidence was not diluted. Instead the dilution effect occurred because the D evidence was very much overestimated, and the D + ND evidence was overestimated to a lesser extent. In this experiment, the D + ND evidence was clearly harder to

evaluate than it was in Experiments 1a and 1b, and participants made more errors as a result. Despite the ease of computing the correct D evidence (i.e., counting the number of matching stimuli in each shop, and taking the ratio), these responses remained on average further from the normative response.

Experiments 1a, 1b, and 2 both used a likelihood ratio scale, which we chose because a simple counting strategy will generate the normative response on this scale. However, as we assessed participant accuracy against the scale used, our conclusions in these experiments depend on participants using the scale correctly. To strengthen our conclusions, it is important to replicate our results using a different response scale.

3. Experiment 3: Responding on a probability scale

In order to determine whether participants in Experiments 1a, 1b, and 2 really were more accurate at evaluating D + ND evidence, or if they are misusing the likelihood ratio scale, we reran Experiment 2 using a different scale: the scale of probabilities. Probability scales have been found to be used correctly in reasoning tasks Fernbach et al. (2010, 2011), Meder et al. (2014), Meder and Mayrhofer (2017) and finding similar results with this scale would strengthen the claim that the dilution effect is not the result of dilution of D + ND evidence.

3.1. Methods

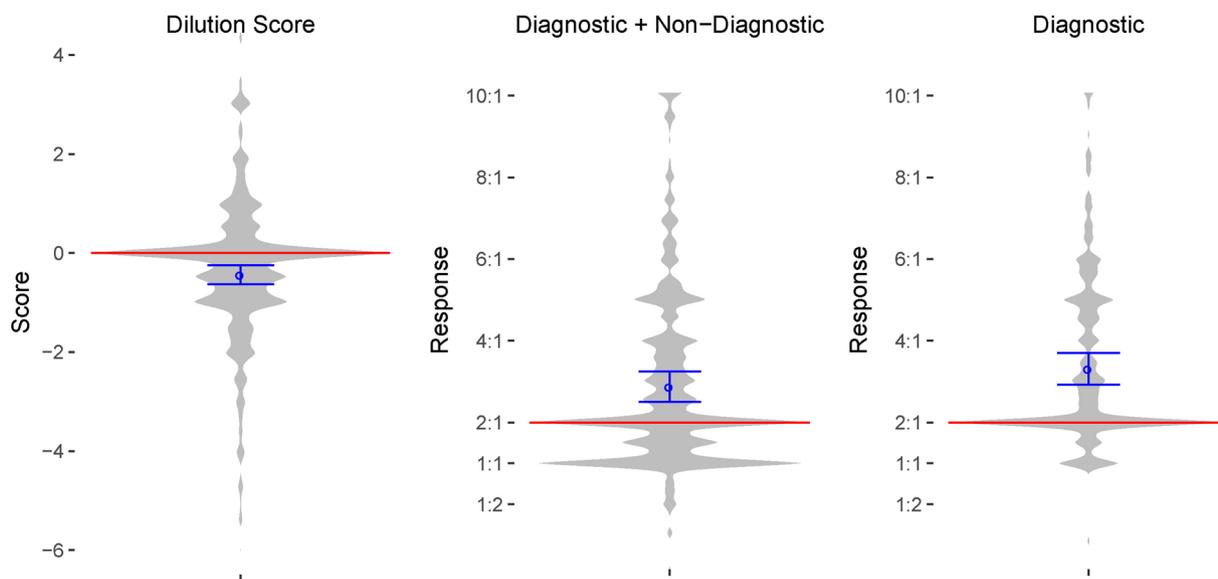
3.1.1. Participants

We recruited 110 participants from Amazon Mechanical Turk, and 105 completed the experiment (60 male, 44 female, and 1 unreported). The average age was 35.2 (SD = 10.7). Participants were paid \$1.50 upon completion of the experiment.

3.1.2. Materials

As in Experiment 2, participants were shown a single pair of ice cream shops, labelled Shop A and Shop B, containing three flavors of ice cream and three flavors of cone in a ratio of 6:3:3 in Shop A and 3:6:3 in Shop B. There were 12 examples of ice cream and 12 examples of cones in each shop, arranged separately instead of combined, as shown in Figure 3. The types and amount of evidence were the same as in Experiment 2.

Participants made responses on a probability scale that ranged from zero to one. The current position of the slider was labelled “From Shop A with a probability of  $X$  and from Shop B with a probability of  $Y$ .” where  $Y$  was equal to  $1 - X$ . If the slider was moved to the extremes of the scale, the label was, “From Shop A with a probability of 1.” or “From Shop B with a probability of 1.” The slider began each trial in the middle of the scale labelled, “Equally probable to be Shop A and B”. Once participants were satisfied with the position of the slider, they



**Fig. 4.** Results of Experiment 2. The means in the Dilution Score plot are equal to the means in the Diagnostic + Non-diagnostic plot minus the means in the Diagnostic plot. Negative values of the dilution score indicate the dilution effect. The red (dark grey) horizontal lines show the normative response in each plot. The blue (dark grey) circles indicate the means across participants, with the error bars showing the 95% confidence intervals of the means. The light grey regions within each plot give Gaussian kernel density estimates (with bandwidth of 0.1) of the of the raw dilution scores and trial-by-trial responses pooled across participants, with the width of each region normalized to its maximum value.

pressed a button to submit their response. The scale was discretized to two decimal points.

3.1.3. Procedure

Participants received similar instructions as in Experiment 2, except they did not receive any instructions to emphasize the independence of how the ice creams and cones were combined. In this experiment, participants responded to 36 stimuli. Each possible D stimulus was repeated 3 times for a total of 12 trials, with an additional 6 trials collecting ND responses. The remaining 18 responses were made to each possible combination of flavors repeated twice each.

The data reported here are from a larger between-participant experiment that tested the effect of alignment of evidence (i.e., separating the ice cream cones so that the ice cream appeared next to the cone), and the way in which the shops and evidence were presented (i.e., presenting all of the information as text rather than as pictures). Here we report the responses of participants in the condition that matched those in Experiments 1a, 1b, and 2.

3.2. Results

As in the experiments above, we calculated a dilution score for each pair of D and D + ND responses, by subtracting the D response from the D + ND response. This score was a difference in probability responses, where a negative value indicated the dilution effect. Figure 5 shows the distribution of these individual dilution scores. While the correct score was most likely, representing 10% of all scores, 78% of errors showed the dilution effect. Using the mean score for each participant, we found evidence for a dilution effect as the dilution score was reliably below zero  $M = -0.065, SD = 0.069, t(104) = -9.64, p < .001, BF_{10} = 1.5 \times 10^{13}$ .

Exactly correct responses (i.e., 2/3) were not possible in this experiment, because the response scale was discretized to two decimal points, but the percentage of D + ND responses above the target value was 53% which was very close to half. The mean D + ND response was very close to and not reliably distinguishable from the normative value of 2/3,  $M = 0.66, SD = 0.072, t(104) = -0.77, p = .44$ , and indeed the Bayes factor showed substantial evidence for the mean being equal to the normative value:  $BF_{10} = 0.144$ . In contrast, the D response were

clearly overestimated:  $M = 0.73, SD = 0.079, t(104) = 7.72, p < .001, BF_{10} = 1.1 \times 10^9$ .

3.3. Discussion

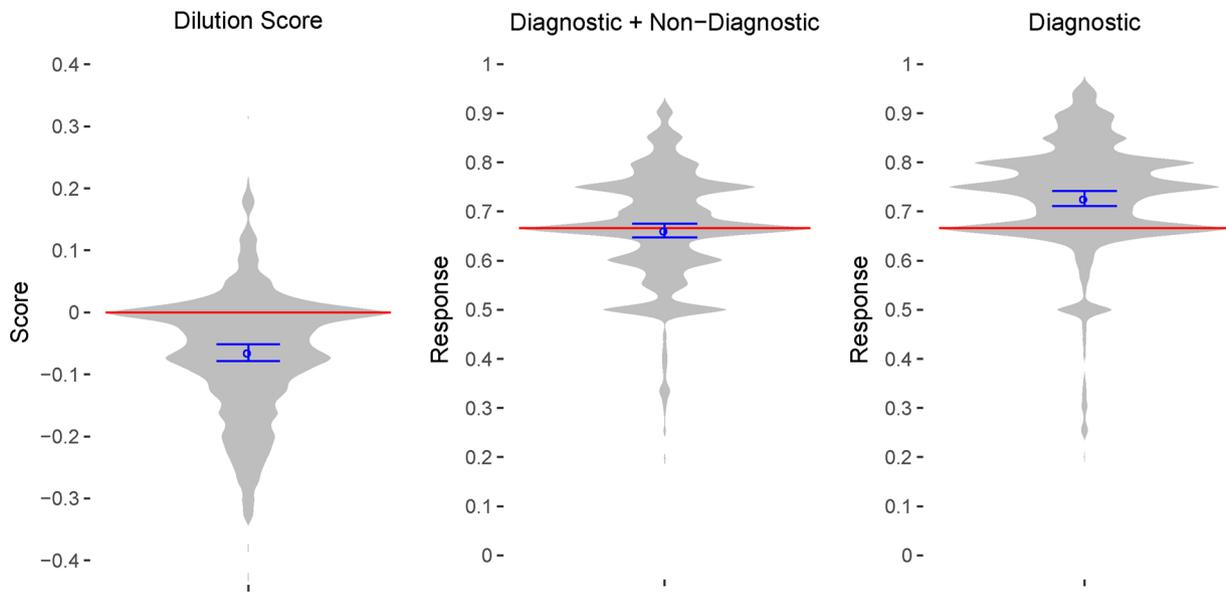
In this experiment we used a probability scale for responses instead of a likelihood ratio scale, again finding a dilution effect without dilution of the D + ND evidence. As in Experiments 1a, 1b, and 2, the dilution effect in this experiment was driven by the overestimation of the D evidence, while the D + ND evidence was on average evaluated more accurately. However, unlike in Experiments 1a, 1b, and 2 which used a likelihood ratio response scale, in this experiment which used a probability scale, participants estimated the D + ND evidence accurately on average.

Extant explanations of the dilution effect do not predict the results from Experiments 1-3. Averaging does not predict our results because averaging assumes that D responses will be unbiased and the D + ND responses would be less than the normative value. Biased hypothesis testing also does not anticipate these results, because it assumes that participants assess each piece of diagnostic evidence as evidence for the favored shop, and additional non-diagnostic evidence as evidence against that shop. Assuming participants are using the scale correctly, biased hypothesis testing should predict that D responses are unbiased, and D + ND responses are underestimated.

Representativeness also does not anticipate these results. Representativeness assumes that participants compare the similarity of the D or D + ND evidence to each of the shops, translating these relative measures of similarity into a response. Representativeness does not make detailed quantitative predictions, so while it has been used to explain the dilution effect, it does not predict whether the D + ND responses or the D responses will be biased. We also provide a more formal investigation in the Appendix showing that a variety of assumptions about similarity do not allow representativeness to match the effects found here.

4. A new explanation for the dilution effect: filling in missing information

Why would participants overestimate the D evidence, and not



**Fig. 5.** Results of Experiment 3. The means in the Dilution Score plot are equal to the means in the Diagnostic + Non-diagnostic plot minus the means in the Diagnostic plot. Negative values of the dilution score indicate the dilution effect. The red (dark grey) horizontal lines show the normative response in each plot. The blue (dark grey) circles indicate the means across participants, with the error bars showing the 95% confidence intervals of the means. The light grey regions within each plot give Gaussian kernel density estimates (with bandwidth of 0.1) of the of the raw dilution scores and trial-by-trial responses pooled across participants, with the width of each region normalized to its maximum value.

underestimate the D+ND evidence? A possibility is that participants are not considering the evidence provided by the ice cream separately from the evidence provided by the cone, but instead are considering ice cream cones as a whole, even when only one part of the ice cream cone is visible. Participants may then be trying to “fill-in” the missing evidence when given only a cone or only an ice cream, analogous to how the brain fills in the blind spot in the retina (e.g., Ramachandran (1992)).

Recent research into the latent scope bias has claimed that it is also the result of people filling in missing features. The latent scope bias occurs when people judge an explanation with fewer unverified predictions as more likely than an explanation with more unverified predictions despite no difference in the diagnostic evidence. In our latent scope bias example from the introduction, the broad-scope explanation was that if the man had murdered his aunt there would be shell casings on the floor and his customary mug of hot chocolate would have been left on the table, and the narrow-scope explanation was that if the second suspect had done it there would only be shell casings. Generally, the unverified predictions are of events with a low base-rate probability (e.g., customary mug of hot chocolate would have been left on the table), but in this recent research a series of experiments showed that the latent scope bias can be reversed by making the unverified predictions of events with high base rates (e.g., if we replaced “customary mug of hot chocolate” with “customary shaker of salt”). The latent scope bias was attributed to participants filling in the missing information according to the base rates of the background information, and then making their judgment using the correct decision rule using this filled-in information as if it were real information Johnson et al. (2016).

We can apply this same explanation to the dilution effect, and filling in from the background distribution can explain the dilution effect we found in Experiments 1a, 1b, and 2. We call this filling-in in an unbiased fashion, to discriminate it from a variant we introduce immediately below. Let's say that a participant is given Shops C and D in Figure 3, and is asked to make a judgment on the likelihood ratio scale about a vanilla ice cream alone. If the missing cone is filled in as chocolate, combined with the vanilla ice cream there would be 4:1 evidence in favor of Shop C, because there are twice as many chocolate cones in

Shop C than in Shop D, and the vanilla ice cream also is twice as prevalent in Shop C: multiplying the 2:1 evidence from the chocolate cone with the 2:1 evidence from the vanilla ice cream produces 4:1 evidence for a vanilla ice cream on a chocolate cone. If the missing cone is filled in as strawberry, combined with the vanilla ice cream there would be 1:1 evidence, because there are twice as many strawberry cones in Shop D than in Shop C which cancels the evidence in favor of Shop C from the vanilla ice cream. If the missing cone is filled in as plain, combined with the vanilla ice cream there would be 2:1 evidence in favor of Shop C, as the plain cone is non-diagnostic. Assuming the participant weights each of these possibilities according to the relative frequency of each cone flavor across both shops (24 in total: 9 chocolate, 9 strawberry, and 6 plain), the participant's response to a vanilla ice cream would then be  $(\# \text{ of chocolate cones in total} / \text{total} \# \text{ of cones}) * (\text{odds that a chocolate cone with vanilla ice cream is in C vs. D}) + (\# \text{ of strawberry cones in total} / \text{total} \# \text{ of cones}) * (\text{odds that a strawberry cone with vanilla ice cream is in C vs. D}) + (\# \text{ of plain cones in total} / \text{total} \# \text{ of cones}) * (\text{odds that a plain cone with vanilla ice cream is in C vs. D}) = (9/24)(4:1) + (9/24)(1:1) + (6/24)(2:1) = 2.375:1$ . In contrast, if the participant is given a plain cone with vanilla ice cream, there are no missing features to fill in, so the participant would just assess the visible stimulus and provide the correct response of 2:1. The dilution effect is produced here because the expected response to the vanilla ice cream (i.e., D stimulus) is greater than the expected response to the plain cone with vanilla ice cream (i.e., D+ND stimulus). The same explanation works in Experiments 1a and 1b.

However, this filling-in mechanism does not generate a dilution effect in Experiment 3, which uses a probability scale. We can simply convert the likelihood ratios in the above example into probabilities, where for likelihood ratio  $x : y$  the probability is  $x / (x + y)$ . This non-linear transformation of the scale has important implications. If the missing cone is filled in as chocolate, combined with a vanilla ice cream the probability response would be 4/5. If the missing cone is filled in as strawberry, combined with a vanilla ice cream the probability response would be 1/2. If the missing cone is filled in as plain, combined with a vanilla ice cream the probability response would be 2/3. So here the response to the vanilla ice cream (i.e., D stimulus) would be  $(9/24)(4/5) + (9/24)(1/2) + (6/24)(2/3) \approx 0.65$ , while the response to the plain

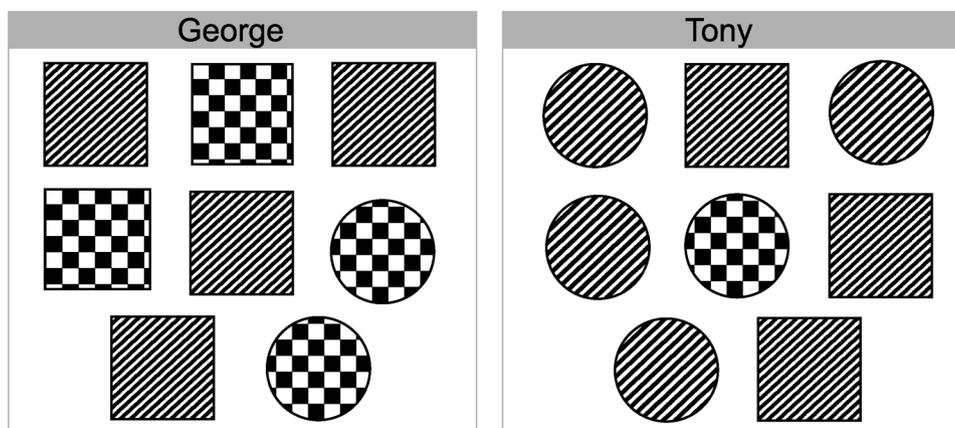


Fig. 6. Example of stimuli used in category-based induction experiments. Each box contains a set of shapes drawn by one of two (fictional) people.

cone plus vanilla ice cream (i.e., D+ND stimulus) would be 2/3. Because the response produced for the D stimulus is less than the response produced by the D + ND stimulus, filling in from the background distribution predicts a slight anti-dilution effect in Experiment 3, which mismatches the dilution effect observed in this experiment<sup>1</sup>.

An alternative is that people may be filling in information in a *biased* fashion. We take as inspiration a close analogue to our task in work on category-based induction. In these studies, participants were presented with sets of shapes with different colors or patterns that were distributed amongst various boxes. Figure 6 shows stimuli similar to those presented in Murphy and Ross (2010). Participants were told that the set of shapes within each box were drawn by a different child, and were then asked about one attribute given another attribute. For example, a participant might be told that the drawing was striped and was then asked what the shape of the drawing would be. The correct answer in this task is to say that the most likely shape is a square, because across the two sets of shapes, there are more striped squares than striped circles.

However, in these category-based induction tasks participants decided upon the missing feature using only the most likely category, instead of looking at both categories. Most participants responded that the most likely shape was a circle, reflecting a process in which participants first identified the most likely set for striped shapes, Tony in Figure 6, and then chose circle because most of Tony's striped shapes were circles Murphy and Ross (1994, 2010).

If people are filling in the missing information using the most likely category, this process might be a key component for explaining the overestimation of the D evidence in all three of the above experiments. When given the vanilla ice cream alone along with the shops in Figure 3, participants may first identify that Shop C is the most likely shop for the vanilla ice cream. Then, looking only at Shop C, they fill in the missing cone with the available flavors using the proportions in that shop. As a result, the mean D response for the vanilla ice cream would be  $(6/12)(4:1) + (3/12)(2:1) + (3/12)(1:1) = 2.75:1$ , which is higher than the 2:1 evidence for D+ND stimuli, producing a dilution effect. Biased filling-in will also produce a dilution effect in Experiment 3 using a probability scale: the predicted response for D is  $(6/12)(0.8) + (3/12)(0.5) + (3/12)(2/3) \approx 0.69$ , which is slightly higher than the

predicted D + ND response of 2/3.

Both types of filling-in are new explanations for the dilution effect, and both types of filling suggest a novel intervention for removing the dilution effect: changing the diagnosticity of what can be used to fill in for the missing feature. Thus, in Experiment 4, we evaluate how changing the evidence available from the missing feature changes the D responses. We investigate whether making the missing feature always non-diagnostic removes the bias in the D responses, and thus reduces or removes the dilution effect.

## 5. Experiment 4: Manipulating the missing evidence

In this experiment, we predict that participant judgments will be influenced by what values are available to fill in for the missing information. If people are filling in missing attributes, in either a biased or unbiased fashion, then their response bias should be reduced when all of the possible values for the missing component are non-diagnostic. Using a likelihood ratio scale, as in the majority of the experiments above, let's assume a participant was asked to judge the relative likelihood of a vanilla ice cream coming from Shops E and F in Figure 7. If either a chocolate or plain cone were substituted for the missing information, this does not change the likelihood ratio: both the chocolate and plain cones are non-diagnostic. The mean response using Shops E and F for filling in for D stimuli would be unbiased,  $(6/12)(2:1) + (6/12)(2:1) = 2:1$ . This forms an interesting contrast with the prediction of a bias in Shops C and D in the Standard condition, despite the two pairs of shops having exactly the same composition of ice creams.

We can also create a pair of shops, Shops G and H, that should strengthen the bias for the D stimuli, as all of the missing information is diagnostic. If a participant were judging the relative likelihood of vanilla ice cream coming from Shops G and H, then two possible cone flavors are always diagnostic. If the participant fills in the most likely cone, chocolate, then the likelihood ratio will be overestimated: 4:1 rather than the correct 2:1. The mean D response from biased filling-in will then be,  $(8/12)(4:1) + (4/12)(1:1) = 3:1$ , which is slightly larger than the predicted mean response of 2.75:1 for Shops C and D in the Standard condition.

### 5.1. Methods

#### 5.1.1. Participants

We recruited 102 participants for this experiment from the University of Warwick community, and 98 participants completed the experiment (81 female and 17 male, with age  $M = 18.8$  and  $SD = 3.3$ ). Participants were randomly assigned to one of three counterbalance groups that determined which pair of shops they saw first. There were 33 participants who saw the Diagnostic Missing Information condition first, 27 participants who saw the Non-diagnostic Missing Information

<sup>1</sup> Taking the weighted average of the evidence along the response scale used in each experiment can be justified by assuming participants believe that they will be evaluated on their squared error along their response scale, as this weighted average minimizes that squared error. As a reviewer suggested, we could instead assume participants always take a weighted average on the probability scale, no matter the scale given in the experiment, as this results in answers that are closer to the correct ones. However, this alternative assumption would result in a worse match between unbiased filling-in and the empirical data.

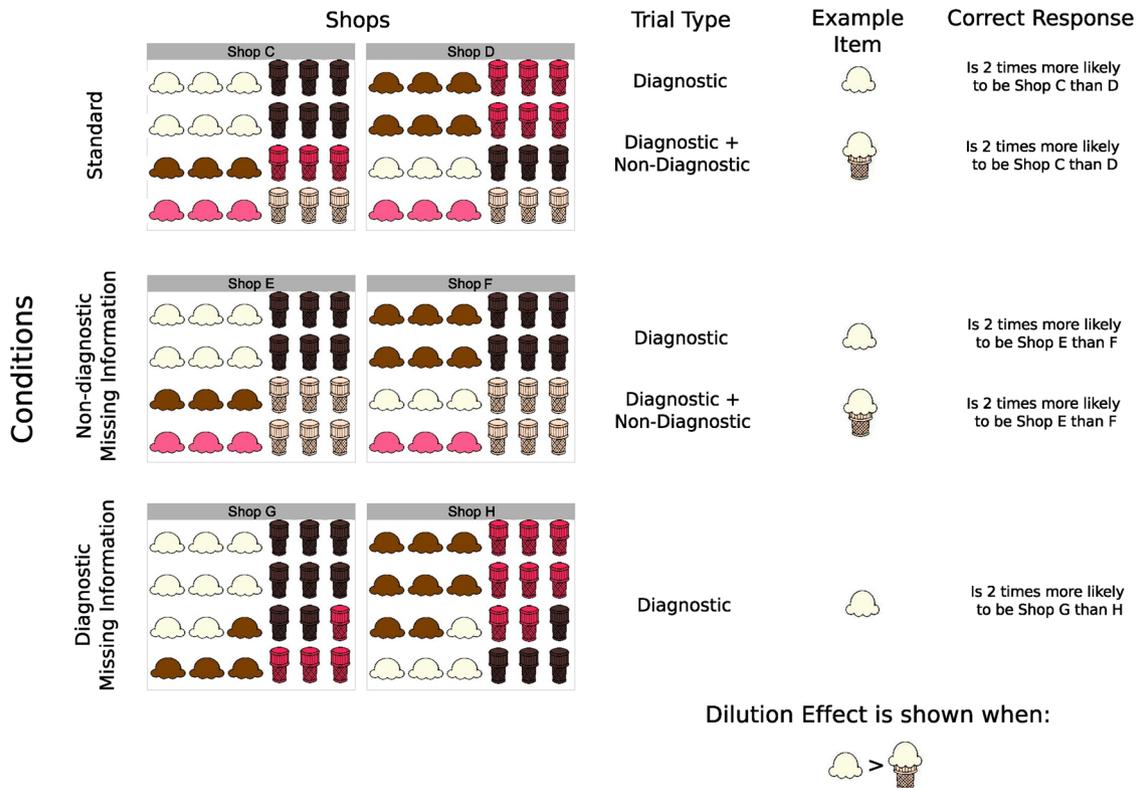


Fig. 7. Example of pairs of shops shown to participants in Experiment 4. Shops C and D are an example of a pair of shops shown in the “Standard” condition, in which the missing information for Diagnostic stimuli can be of any type. Shops E and F are an example of a pair of shops shown in the “Non-diagnostic Missing Information” condition, in which the missing information for Diagnostic stimuli is always non-diagnostic. Shops G and H are an example of a pair of shops shown in the “Diagnostic Missing Information” condition, in which the missing information for Diagnostic stimuli is always diagnostic.

condition first participants, and 38 participants who saw the Standard condition first.

5.1.2. Materials

Participants were given three different types of shop pairs in the experiment. All shop pairs contained 12 examples of ice creams and 12 examples of cones, arranged separately, as in Experiment 2. One type of shop pair, termed Standard, had the same structure as all of the shop pairs in Experiment 2 as exemplified by Shops C and D in Figure 3. For both ice creams and cones, there was one flavor that had 2:1 diagnostic evidence for one shop, another flavor with 2:1 diagnostic evidence for the other shop, and the third flavor was non-diagnostic. There were 15 possible test stimuli for this shop pair: 6 ice creams or cones alone, and 9 combinations of ice cream and cone.

The second shop pair, termed Non-diagnostic Missing Information with an example given by Shops E and F in Figure 7 did not have the same symmetry as in the Standard pair. While one component, either cone or ice cream had the same 6:3:3 and 3:6:3 ratios in the two shops, the other component was always non-diagnostic, having the ratios 6:6 and 6:6 in the two shops. This means that for D stimuli, the missing information is always non-diagnostic in this condition. There were 11 possible test stimuli for this shop pair: 5 ice creams or cones alone, and 6 combinations of ice cream and cone.

The third shop pair, termed Diagnostic Missing Information, did not contain any non-diagnostic evidence, as exemplified by Shops G and H in Figure 7. Instead, both ice creams and cones always gave diagnostic evidence at the 2:1 level. This means that for D stimuli, the missing information is always diagnostic in this condition. The ratios of ice creams and cones were always 2:1 for one shop, and 1:2 for the other shop. There were 8 possible test stimuli for this shop pair: 4 ice creams or cones alone, and 4 combinations of ice cream and cone.

The response scale used was the same as in Experiments 1a, 1b, and

2. We corrected for the programming error in the scale in the same way it was corrected for in the earlier experiments, which affected 6.0% of responses in this experiment.

5.1.3. Procedure

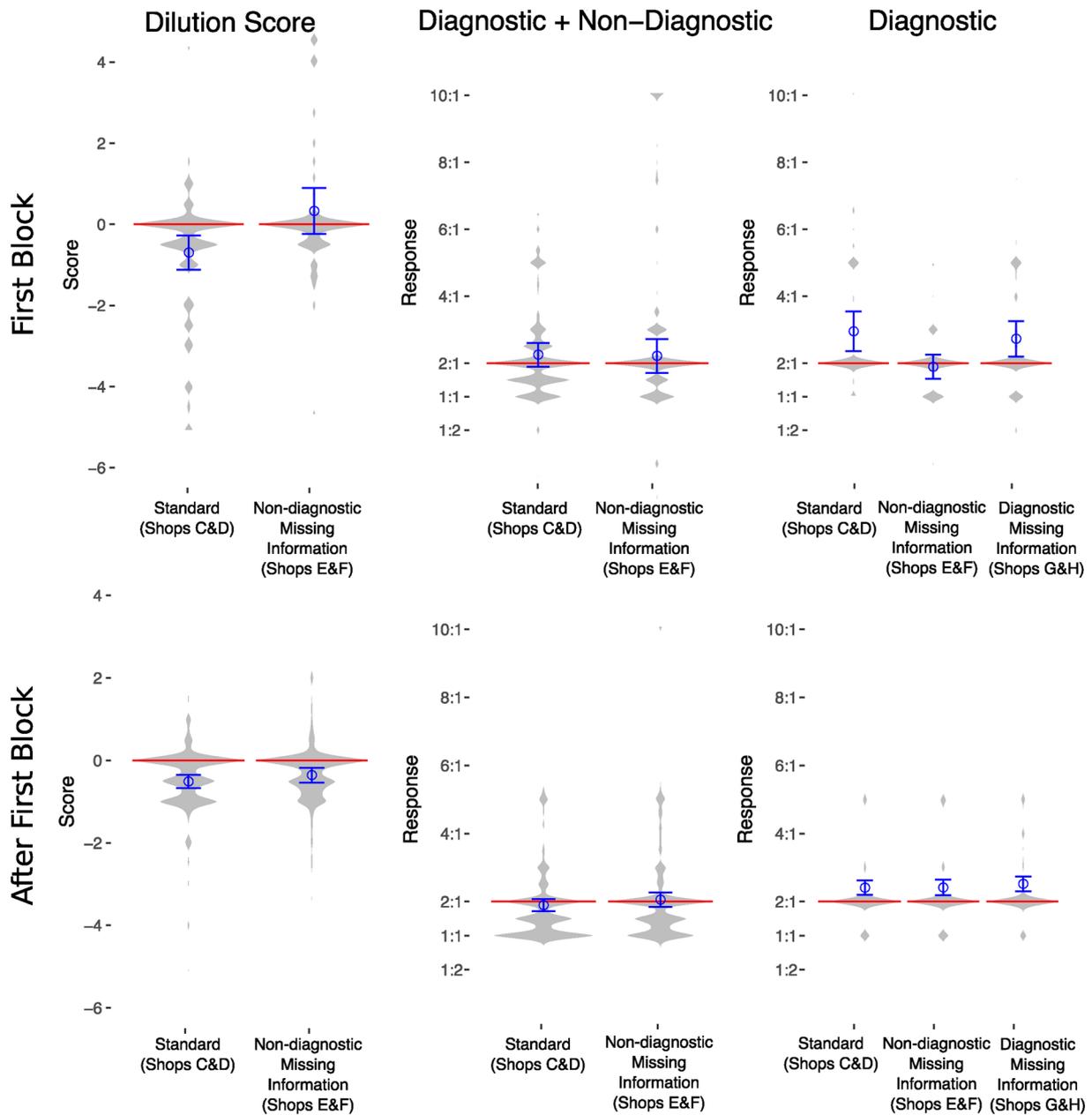
Participants saw all three conditions twice each, with the order of conditions randomized so that participants saw all three conditions before they repeated a condition. Within a condition, participants saw a pair of shops that fit that condition, and made judgments about all possible ice creams and cones alone and all possible pairs of ice creams and cones that could arise from that pair of shops. As a result, there were different numbers of trials in each block: 15 trials for a block in the Standard condition, 11 trials for a block in the Non-diagnostic Missing Information condition, and 8 trials for a block in the Diagnostic Missing Information condition.

The instructions were similar to those used in Experiments 1a, 1b, and 2. Between pairs of shops, participants were told that they were “moving on to a new town” and that in that new town they would make the same kinds of judgments but the shops would be different and have different names. Also to encourage participants to treat the ice cream and cones in the shops as being selected independently, they were told, “People like to mix and match ice cream and cones, and buy every combination of ice cream and cones.”

5.2. Results

5.2.1. Results of the first block

We first report the results of the first pair of shops seen by participants, because we found that responses changed after the first block. As in the experiments above, we calculated a dilution score for each pair of D and D + ND responses, by subtracting the D response from the D + ND response. This score was a difference in likelihood ratio responses,



**Fig. 8.** Results of Experiment 4. The upper row of plots presents data from the first block of the experiment, and the bottom row presents the data aggregated over the remaining blocks. The means in the Dilution Score plot are equal to the means in the Diagnostic + Non-diagnostic plot minus the means in the Diagnostic plot. Negative values of the dilution score indicate the dilution effect. Along the horizontal axis of each plot are the condition labels, with the shops from Figure 7 relevant to that condition in parentheses. The red (dark grey) horizontal lines show the normative response in each plot. The blue (dark grey) circles indicate the means across participants, with the error bars showing the 95% confidence intervals of the means. The light grey regions within each plot give Gaussian kernel density estimates (with bandwidth of 0.1) of the of the raw dilution scores and trial-by-trial responses pooled across participants, with the width of each region normalized to its maximum value.

where a negative value indicated the dilution effect. Figure 8 shows these individual dilution scores. Note that this can only be done for the Non-diagnostic Missing Information and the Standard conditions, because the Diagnostic Missing Information condition did not have any D + ND responses that could be used to calculate dilution scores. All tests of mean differences in responses between participants were performed with Welch's unequal variances t-test, as this test is robust to differences in standard deviation or sample sizes.

We found reliably less dilution in the Non-Diagnostic Missing Information condition compared to the Standard condition,  $t(52.1) = -2.97, p = .004, BF_{10} = 10.8$  (means and standard deviations reported below). This difference between the Non-diagnostic Missing Information and Standard conditions was due to differences in the D

responses because these differed between conditions,  $t(57.9) = -3.12, p = .003, BF_{10} = 6.6$ , while there was no difference in the D + ND responses,  $t(49.9) = -0.12, p = .91, BF_{10} = 0.26$ . The D responses also differed between the Non-diagnostic Missing Information and Diagnostic Missing Information conditions,  $t(54.0) = -2.66, p = .010, BF_{10} = 3.6$ , though they did not differ between the Standard and Diagnostic Missing Information conditions,  $t(68.9) = 0.58, p = .567, BF_{10} = 0.28$ .

Looking at the individual conditions in more detail, in the Standard condition, the correct score was most likely, representing 34% of all scores, and of the errors, 77% showed the dilution effect. Using the mean score for each participant, we found evidence for a dilution effect as the dilution score was reliably below zero,  $M = -0.70, SD = 1.28$ ,

$t(37) = -3.37, p = .002, BF_{10} = 18.6$ . D responses in the Standard condition were correct on 57% of trials, and participants reliably overestimated the evidence,  $M = 2.95, SD = 1.80, t(37) = 3.26, p = .002, BF_{10} = 14.2$ . The D+ND responses in the Standard condition were correct on 30% of trials, and were not reliably different from the normative response,  $M = 2.25, SD = 1.07, t(37) = 1.44, p = .16, BF_{10} = 0.45$ .

By contrast the dilution errors in the Non-diagnostic Missing Information condition were more balanced: the correct score was again most likely, representing 46% of all scores, and of the errors 59% showed the dilution effect. Using the mean score for each participant, we found no evidence for a dilution effect as the mean dilution score was actually above zero,  $M = 0.33, SD = 1.43, t(26) = 1.18, p = .25, BF_{10} = 0.38$ . D responses in the Non-diagnostic Missing Information condition were correct on 61% of trials, and participants did not overestimate the evidence,  $M = 1.89, SD = 0.91, t(26) = -0.63, p = .53, BF_{10} = 0.24$ . The D+ND responses in the Non-diagnostic Missing Information condition were correct on 39% of trials, and were also not overestimates of the evidence,  $M = 2.21, SD = 1.27, t(26) = 0.88, p = .39, BF_{10} = 0.29$ .

It was not possible to calculate dilution scores in the Diagnostic Missing Information condition because no D+ND responses were possible. This condition was used as a control to check whether any effects found in the Non-diagnostic Missing Information condition were due only to using two flavors instead of three. D responses in the Non-diagnostic Missing Information condition were correct on 48% of trials. Unlike in the Non-diagnostic Missing Information condition, participants did overestimate the evidence in this condition,  $M = 2.72, SD = 1.50, t(32) = 2.78, p = .009, BF_{10} = 4.75$ .

### 5.2.2. Results of later blocks

Unlike in our previous experiments, in this experiment participants experienced multiple conditions. We found that the dilution scores increased for the Non-diagnostic Missing Information condition in later blocks<sup>2</sup>, as shown in Figure 8,  $t(34.2) = 2.84, p = .008, BF_{10} = 35.7$ . This appears due to an increase in the D responses,  $t(63.8) = -3.14, p = .003, BF_{10} = 5.73$ , as the D+ND responses did not increase,  $t(43.9) = 0.42, p = .68, BF_{10} = 0.25$ , in this condition. In contrast, the dilution scores in the Standard condition did not reliably change between the first and later blocks,  $t(52.3) = -1.16, p = .25, BF_{10} = 0.46$ . However, both the D,  $t(48.5) = 2.35, p = .022, BF_{10} = 5.12$ , and D+ND responses,  $t(70.1) = 2.24, p = .029, BF_{10} = 2.23$ , appeared to shift lower in the Standard condition after the first block.

Comparing the dilution effect in the Non-diagnostic Missing Information condition the second time this condition was experienced, we also found that participants who experienced this condition in their first block showed a reduced dilution effect compared to participants who experienced another condition in their first block,  $t(83.0) = 2.02, p = .046, BF_{10} = 0.69$ , but as the t-test and Bayes factor disagree on the direction of the effect, the evidence is weak.

Unlike in the first block, in the remaining blocks we found no difference in the dilution scores in the Non-Diagnostic Missing Information condition compared to the Standard condition,  $t(193.8) = -0.46, p = .65, BF_{10} = 0.17$ . There was also no reliable difference between conditions in the D responses,  $t(193.0) = 0.64, p = .52, BF_{10} = 0.19$  or the D+ND responses,  $t(191.1) = 1.09, p = .29, BF_{10} = 0.27$ . Additionally there were no reliable difference in D responses between the Non-diagnostic Missing Information and Diagnostic Missing Information conditions,  $t(193.1) = -0.619, p = .537,$

<sup>2</sup>To compare statistically across blocks, we divided participants into those who saw a particular condition in their first block and those who did not and performed a between-participants comparison. For participants who did not have a particular condition in their first block, we aggregated responses over the two repetitions of the block. In Figure 8 we show all the data.

$BF_{10} = 0.19$ , or between the Standard and Diagnostic Missing Information conditions,  $t(194.0) = -1.30, p = .194, BF_{10} = 0.34$ .

Looking at the individual conditions in more detail, in the Standard condition the dilution score was reliably below zero  $M = -0.49, SD = 0.85, t(97) = -5.67, p < .001, BF_{10} = 87136$ . D responses were reliably overestimated compared to the normative response,  $M = 2.32, SD = 1.05, t(97) = 2.98, p = .003, BF_{10} = 7.1$ . The D+ND responses were not reliably different from the normative response,  $M = 1.83, SD = 0.93, t(97) = -1.84, p = .07, BF_{10} = 0.56$ . The Non-diagnostic Missing Information now showed a reliable dilution effect,  $M = -0.43, SD = 0.82, t(97) = -5.21, p < .001, BF_{10} = 13202$ . D responses in the Non-diagnostic Missing Information condition were now reliably above the normative value,  $M = 2.41, SD = 1.13, t(97) = 3.64, p < .001, BF_{10} = 47.5$ . The D+ND responses however were still not reliably different from the normative response,  $M = 1.98, SD = 1.05, t(97) = -0.18, p = .86, BF_{10} = 0.11$ . In the Diagnostic Missing Information condition, participants did again overestimate the D evidence,  $M = 2.51, SD = 1.05, t(97) = 4.81, p < .001, BF_{10} = 2754$ .

### 5.3. Discussion

In the first block of this experiment, we saw the effect we expected assuming participants were filling in the missing information. In both the Standard and Diagnostic Missing Information conditions, the most likely shops contained a plurality of diagnostic flavors for the missing attribute and D evidence was greatly overestimated. In the Non-diagnostic Missing Information condition, the flavors of the missing attribute for the most likely shop were all non-diagnostic, and there was no dilution effect in this condition. This experiment strengthens the qualitative evidence for filling-in because it does not require participants to have used the scale correctly: participants overstating their evaluation of each piece of evidence (i.e., shifting their responses in each condition the same amount) could explain the results in Experiments 1-3, but not in this experiment. In addition, this experiment shows a reliable effect of the distribution of evidence in the *missing* feature on the responses to D stimuli. One alternative explanation is that participants may realize that the missing feature in the Non-diagnostic Missing Information condition is non-informative, and therefore not fill it in, and while we cannot exclude this possibility, it would still implicate filling-in as the source of the dilution effect.

Once participants had participated in the other conditions however, the Non-diagnostic Missing Information condition was no longer effective in removing the dilution effect, and the D responses were the same as in the Standard condition. The lack of difference between conditions is interesting. It could be that the bias away from the normative responses were still mainly due to filling in the missing information, and that participants were reusing the D responses across blocks. In particular, participants may have seen that the evidence favored one shop over the other and rather than recalculate their response, they just reused the response they made to similar questions in a previous block. Indeed, participants who saw the Non-diagnostic Missing Information condition first showed a smaller dilution effect when this condition came up a second time than participants who saw another condition first, though this effect was weak. This kind of computation-saving strategy has been called amortized inference Gershman and Goodman (2014) and could potentially explain the differences between responses in our first and later blocks.

## 6. General discussion

Across a series of experiments, we found that participants did show a dilution effect for simple objective stimuli. Surprisingly, however, we found that in most conditions participants were more accurate when making judgments of D+ND evidence than they were when judging D evidence alone. This observation goes against the very name of the

dilution effect: instead of ND evidence diluting D evidence, it appears that participants had trouble judging D evidence on its own. The result was reliably found for different evidence strengths, held even when it was more difficult to assess the D+ND evidence than the D evidence alone, and occurred with both likelihood ratio and probability response scales. These results were not anticipated by past explanations of the dilution effect including representativeness [Kahneman and Tversky \(1972\)](#), [Nisbett et al. \(1981\)](#), averaging [Anderson \(1967\)](#), [Shanteau \(1975\)](#), or biased hypothesis testing [Meyvis and Janiszewski \(2002\)](#).

We proposed a new explanation for the dilution effect: participants see the D evidence as having a missing feature and filling in that missing feature using the information available in the most likely hypothesis, which causes an overestimate of the D evidence. Filling-in has been used to explain the latent scope bias in past work [Johnson et al. \(2016\)](#), assuming that participants fill in missing information using the background distribution of possibilities. We modified this hypothesis to explain the dilution effect, proposing that assumes participants fill in the missing information using the most likely hypothesis, similar to that which has been found in category-based induction [Murphy and Ross \(1994, 2010\)](#). This biased filling-in process could explain the dilution effect observed in Experiment 3, which used a probability scale, as well as in the other experiments. We found further support for biased filling-in in a quantitative model comparison reported in the Supplementary Material.

The filling-in hypothesis also allowed us to identify a novel manipulation for eliminating the dilution effect: making the values available to fill in for the missing information entirely non-diagnostic. We found that, at least initially, this made the dilution effect disappear as the mean D response became very close to the normative response.

In the below discussion, we explore what kinds of dilution effects the filling-in hypothesis can explain, what other kinds of dilution effects it cannot explain, and hypothesize about what would drive participants to use biased filling-in as a judgment strategy.

### 6.1. Dilution effects that filling-in can explain

Filling-in can potentially explain how various manipulations used by other researchers affect the dilution effect. Filling-in requires more steps to calculate than the correct answer does, as we argue below, which may explain why time pressure reduces the dilution effect for audit decisions [Glover \(1997\)](#). In addition, if filling-in is a misunderstanding of the task rather than a heuristic used to reduce effort, then this may explain why the most expert participants show a reduction in the dilution effect [Shelton \(1999\)](#).

Filling-in may also explain one key piece of evidence [Meyvis and Janiszewski \(2002\)](#) used to motivate biased hypothesis testing as an explanation for the dilution effect. In their Experiment 2, they used not just non-diagnostic evidence, but also weakly diagnostic evidence in combination with diagnostic evidence (D+WD). While a dilution effect was found as mean D responses were greater than mean D+ND responses, mean D responses were less than mean D+WD responses. This result was key evidence against averaging, because averaging predicts that D+WD responses would be less than D responses. However, this result can potentially be explained by filling-in. If D+ND and D+WD responses are estimated accurately, then D+WD will exceed D+ND, as the empirical data show. If the only possibilities for the missing feature are ND or WD, which for the subjective stimuli of [Meyvis and Janiszewski \(2002\)](#) would require empirical work to verify, then filling in using either ND or WD evidence would result in D responses that fell between the two, replicating the empirical pattern.

Filling-in can perhaps explain the dilution effect found in the perceptual data of [Hotaling et al. \(2015\)](#), which asked participants to judge which of two real faces a face stimulus resembled more. The face stimuli in this experiment were morphs between the top halves of the two real faces, morphs between the bottom halves of the two faces, or both. When both were presented, this was done in two different ways: one in

which the two halves of the face were offset so they appeared “split”, and one in which the two halves were aligned. Participants were not given non-diagnostic evidence, but were instead given diagnostic evidence of various strengths for each face, and the dependent variable was participant accuracy rather than participant ratings of evidence strength. The empirical result was that larger difference in evidence strength led to a larger dilution effect.

A key difference between our task and that of [Hotaling et al. \(2015\)](#) is that they assessed accuracy in a task in which it was more difficult to know which face was the most likely source of each evidence, compared to our task in which careful counting can perfectly determine the most likely source of each piece of evidence. Whether filling-in can explain the results depends on how participants are making their choice: selecting the response probabilistically, or always selecting what they consider to be the most likely response [Acuna et al. \(2015\)](#), [Acerbi et al. \(2014\)](#), [Drugowitsch et al. \(2016\)](#), [Sanborn and Beierholm \(2016\)](#). Filling-in could explain the overall dilution effect in this experiment assuming that participants are constructing a subjective probability distribution that each response is correct, and then making a response according to those probabilities, a decision strategy called probability matching. In that case, filling in the missing evidence using the most likely face when only a half-face is shown would raise the subjective probability of the face the participant considered more likely – it would reduce the amount of probability matching that was done and hence raise the accuracy of the half faces. However, if participants are not probability matching, and instead are responding by always selecting the stimulus they consider subjectively more likely, then filling-in would raise confidence and but not change their decision, and the dilution effect would need a different explanation in this experiment. And indeed, it will require detailed computational modelling to see if all of the results could be matched by a biased filling-in account.

One key aspect of the ice cream cone stimuli used in all of our tasks was that it was clear that there should be two pieces of evidence: a single cone or a single ice cream signaled that there was missing evidence. Many other kinds of stimuli do not have this clear signaling of missing evidence, which becomes especially important when the dilution effect is obtained between-participants, by some participants judging D stimuli and others judging the D+ND stimuli. Between-participant designs in which it is not obvious that there is missing evidence constitute many of the experiments on the dilution effect, including those asking for judgments or predictions about people [Kemmelmeyer \(2004, 2007\)](#), and those asking for consumers’ judgments of products [Igou and Bless \(2005\)](#).

Despite this lack of clear signaling, it is possible that filling-in still plays a role. Returning to the example from [Zukier and Jennings \(1984\)](#) given at the beginning of this paper, mock jurors were more convinced that a man had murdered his aunt with a handgun when they were told diagnostic information, than when they were told the diagnostic information plus the non-diagnostic information that he was of average height and had average vision. When imagining whether a person had committed a murder, it seems reasonable to consider whether this person is physically capable of the act. It may well be that participants are filling in physical information about the man in a biased fashion: imagining him as large and able to accurately aim a gun. Relative to this, the non-diagnostic information of being of average height and vision could make seem less capable and thus less likely to be guilty. Congruent to this, [Zukier and Jennings \(1984\)](#) ran a third condition in which the non-diagnostic information was replaced with atypical non-diagnostic information, that the man was extremely tall and had very good vision, and found that impressions of guilt in this condition did not significantly differ from that when participants were given diagnostic information alone. If participants were filling in similar information for the missing physical attributes, this could explain why no difference was found. Overall, filling could operate where there is missing information that is *potentially* diagnostic.

## 6.2. Multiple causes of the dilution effect

While filling-in can explain some aspects of dilution effect, there are other aspects of the dilution effect that seem to require a different explanation. Much research in the dilution effect has manipulated exactly what kind of non-diagnostic information is given to participants [Fein and Hilton \(1992\)](#), [Hilton and Fein \(1989\)](#), [Peters and Rothbart \(2000\)](#). [Peters and Rothbart \(2000\)](#) suggested that these results occur because participants use non-diagnostic information to disambiguate diagnostic information. For example, when making predictions about a person A's criminal behavior, finding out "person A is an alcoholic" suggests a range of bad outcomes, but it is not clear from that statement how severe the alcoholism is. Non-diagnostic information that person A manages a hardware store helps show that the alcoholism is not as bad as it could be, providing useful indirect information that should reduce the prediction of criminal behavior. This explanation supposes that participants are correctly evaluating both D and D+ND evidence, but the experimenter is making the incorrect assumption that the evidence is independent. This normative form of evidence combination of course needs no filling-in process to explain, though it is consistent with the idea behind filling-in that participants are not using simple rules, but are engaging in complex inference when making judgments in these tasks.

While in the preceding section, we argued that at times people may fill-in even in between-participant experiments, there may well be other causes to the dilution effect found in these experiments. Many of these experiments have investigated whether the dilution effect may be due to Gricean conversational norms. The idea is that participants believe that the experimenter would not give them irrelevant evidence because people generally do not insert irrelevant evidence into conversations, so the participants treat the non-diagnostic as diagnostic. This idea has found support in some work [Igou and Bless \(2005\)](#), [Igou \(2007\)](#), though other work has shown that a conversational bias produces no dilution effect on its own [Kemmelmeyer \(2007a,b\)](#). Instead, it seems that participants need to be both convinced that the conversational bias does not apply and be held accountable for their responses in order for the dilution effect to disappear [Tetlock et al. \(1996\)](#).

## 6.3. Biased filling-in as a computational strategy

We explained the dilution effect in our experiments with a mechanism that fills in missing values in a biased fashion. The most interesting aspect of filling-in is that it occurs at all. As we point out in the introduction to Experiments 1a and 1b, there is a simple and correct algorithm for implementing the normative model for D stimuli: count the number of matching stimuli in each shop (e.g.,  $x$  and  $y$ ) and report the ratio (e.g.,  $x/y$ ) on the likelihood ratio scale, or the relative frequency ( $x/(x + y)$ ) on the probability scale. Compared to this normative algorithm, the filling-in process requires additional processing steps for D stimuli: first filling in the missing information, and next calculating the answer for the more complex filled-in stimulus.

Because it involves additional steps to calculate compared to the normative algorithm, filling-in does not save any time or effort on the part of the participant. This contrasts filling-in with other common explanations of reasoning biases like representativeness or biased hypothesis testing, which assume that participants substitute an easier-to-calculate answer for a difficult answer [Kahneman and Tversky \(1972\)](#), or averaging, which assumes that participants are using a cognitively simpler and more robust strategy [Juslin et al. \(2009\)](#). As filling-in instead assumes that participants are going out of their way to make an incorrect response, it suggests that it might be a useful strategy in other tasks.

So why would filling-in be a useful strategy? We can look for inspiration to the statistics literature, where filling-in describes a set of methods collectively known as *imputation* (for an overview, see [Little and Rubin \(2014\)](#)). Imputation is the process of filling in missing values

in a data set and is often used in these situations. It can be particularly useful when data are not missing at random, but instead missing due to a known cause. For example, assuming the distribution of p-values observed in journal articles is representative of the p-values obtained in experiments would be a mistake because lower p-values are more likely to result in publication. Methods such as trim-and-fill have been used to correct for this bias by imputing the missing p-values [Duval and Tweedie \(2000\)](#).

In everyday experience these kinds of selection biases are rife in the information that people observe: people choose their news sources and their friends, both of which are likely to result in biased information about the world (e.g., [Del Vicario et al. \(2016\)](#)). Imputation, when done properly, can help correct for these kinds of biases. Some laboratory experiments have shown that people do use imputation to correct for biased missing data when they themselves are the ones who choose which data to observe [Denrell et al. \(2019\)](#), [Elwin et al. \(2007\)](#), [Henriksson et al. \(2010\)](#). Filling-in is then perhaps useful as a default strategy to correct for the biased missing data that results from motivated sampling of information.

While filling-in may be a useful strategy, the imputed values in both our experiments and the latent scope bias experiments are biased. In the work on filling-in in the latent scope bias, errors in reasoning with missing evidence occurred because participants were inferring the missing evidence from the background base rates, rather than inferring it from the hypotheses under consideration [Johnson et al. \(2016\)](#), [Johnston et al. \(2017\)](#). In our experiments, participants appear to have preferentially inferred the missing evidence from the hypothesis most likely to have produced the observable evidence, in line with work on category-based induction [Murphy and Ross \(2010\)](#). Perhaps these biased values are used to save time or effort over generating better imputations, as these works have hypothesized.

Of course even biased filling-in assumes is a rather complex process which averages over the many possible ways of filling in the missing evidence. This process is not necessarily what participants are doing. A computationally simpler alternative, motivated by the idea that people approximate complex inference through sampling [Griffiths et al. \(2012\)](#), [Sanborn et al. \(2010\)](#), [Sanborn and Chater \(2016\)](#), is that participants are implicitly sampling the value that is filled in rather than averaging over all of the possibilities. That is, they are filling in a single value for the missing feature, which is picked according to the probability it could occur. Averaging over a number of different judgments in which a single missing value is sampled will produce equivalent predictions to the biased filling-in account we describe above, and our above analyses do not discriminate between these possibilities.

Looking at the distributions of individual responses however does suggest that people are taking one sample (or a small number of samples) of the missing evidence: filling in with a single sample would produce a "spiky" response distribution with spikes at each of the levels of evidence that filling-in could produce. For example, in Experiment 2, filling in with a single sample would produce responses of 1:1, 2:1 or 4:1. Of course this is not the only explanation of spiky response distributions – our justification for analyzing the mean responses was that participants often round their responses. Determining the extent to which people are sampling, rounding, or both will require careful work to disentangle these processes.

## 6.4. Conclusion

Overall, it is quite likely that there are multiple causes of the dilution effect, each which may be more or less to blame for the effect depending on the design of the experiment. Our contribution is to demonstrate that a completely different type of hypothesis, filling-in, is also needed to account for the dilution effect. Unlike other explanations, filling-in explains the dilution effect as a misestimation of the diagnostic evidence alone, rather than a dilution of the diagnostic

evidence by non-diagnostic evidence.

Experimentally, this work points to the need to be careful about assuming where the error is when evaluating human cognition against normative models. What experimenters consider to be the judgments that are easy and unbiased may not necessarily be the case. Indeed, there is the potential for filling-in to explain, at least in part, additional decision making biases. For example, the effects of conservatism in probability judgments, the conjunction and disjunction fallacies, and the effect of subadditivity all require participants to judge both parts and wholes, and the biases are demonstrations that the judgment of the part is higher than it should be relative to the judgment of the whole. While these effects have all been explained as a result of inappropriate aggregation of evidence, filling-in may also be to blame.

## Appendix A. Appendix

In this Appendix, we explore the predictions of representativeness. While representativeness has often appealed to similarity as defined in Tversky's contrast model Nisbett et al. (1981), Tversky (1977), our stimuli consisted of shops that displayed examples in different proportions, pointing more naturally to similarity as conceptualized in exemplar models of categorization. We show here, that depending upon the assumptions made, that for the predictions of the  $D$  and  $D + ND$  stimuli, representativeness will agree with either the normative model, the averaging model, or look like indifference.

Exemplar similarity depends on the mismatch between features, which we assume to be zero for matching ice creams or matching cones, and equal to  $d$  for any mismatches. Similarity also depends on the exponent  $r$ , which we assume to be equal to 1 because the features are separable. Finally, similarity also depends on the generalization gradient,  $c$ , which in our example here is confounded with  $d$ , so in the below we assume  $c = 1$  and just investigate  $d$ . There are two main ways in which similarity can be combined across features. The most common assumption is that similarity across features is multiplied, so the similarity between a stimulus and a single item in a shop is  $\frac{1}{m}$ , where  $m$  is the number of mismatches. For simplicity, we set  $s = e^{-d}$  in the below, where  $0 \leq s \leq 1$  is the decrement in similarity for one mismatch. To find the total similarity between an item and a shop, the individual similarities between a stimulus and the shop items is summed Nosofsky (1986). Finally for a response on the likelihood ratio scale, there is little in past research to guide us. We make the straightforward assumption that people take the ratio of similarities, and for a response on a probability scale, we assume that people take the normalized ratio of similarities, as is assumed for categorization judgments.

For multiplicative similarity, the predicted similarity ratios between  $D$  and  $D + ND$  stimuli and Shops A and B ( $S_A$  and  $S_B$  respectively) in Experiment 1a is the same, no matter how similar one mismatch is

$$\frac{S(D, S_A)}{S(D, S_B)} = \frac{8 + 8s}{4 + 12s} = \frac{2 + 2s}{1 + 3s} \quad (1)$$

$$\frac{S(D + ND, S_A)}{S(D + ND, S_B)} = \frac{2 + 8s + 6s^2}{1 + 6s + 9s^2} = \frac{(1 + 3s)(2 + 2s)}{(1 + 3s)^2} = \frac{2 + 2s}{1 + 3s} \quad (2)$$

Depending on  $s$ , the predictions for  $D$  and  $D + ND$  stimuli range from the normative values of 2:1 if  $s = 0$  to indifference of 1:1 if  $s = 1$ . For this assumption of multiplicative exemplar similarity, representativeness can vary between the normative model and indifference.

Alternatively, it could be that participants are using additive similarity. Additive similarity has been discredited in categorization experiments because it does not allow exemplar models to match the representational flexibility of human behavior Nosofsky (1992), but it is possible that it might be used in Experiments 2-4 because the features are presented separately in the shops. If additive similarity is used, then predicted similarity ratios between  $D$  and  $D + ND$  stimuli and Shops C and D ( $S_C$  and  $S_D$  respectively) in Experiment 2 are

$$\frac{S(D, S_C)}{S(D, S_D)} = \frac{6 + 18s}{3 + 21s} = \frac{2 + 6s}{1 + 7s} \quad (3)$$

$$\frac{S(D + ND, S_C)}{S(D + ND, S_D)} = \frac{9s + 15s^2}{6s + 18s^2} = \frac{3 + 5s}{2 + 6s} \quad (4)$$

Depending on  $s$ , like for multiplicative similarity, the predictions for  $D$  judgments range from the normative values of 2:1 if  $s = 0$  to indifference of 1:1 if  $s = 1$ . However, for additive similarity, the predictions for  $D + ND$  judgments are different, ranging from 1.5:1 if  $s = 0$  to indifference of 1:1 if  $s = 1$ . Thus if  $s < 1$ , additive similarity predicts a dilution effect. For this assumption of additive exemplar similarity, representativeness can thus vary between the averaging model and indifference.

None of the possibilities considered here for exemplar similarity predict that  $D$  stimuli are judged to be greater than the normative value of 2:1, as observed in our empirical data. Additionally, none of the possibilities predict that judgments of  $D$  stimuli are influenced by what is displayed in a shop for the missing feature, as shown in the first block of Experiment 4. However, we should be clear that we are investigating exemplar similarity within a pure representativeness explanation, and not making claims about exemplar similarity more generally. Indeed, exemplar models have been modified via a filling-in process similar for category labels to produce the results of category-based induction tasks Nosofsky (2015), and biased filling-in with an exemplar model could likely account for our results as well.

## References

- Acerbi, L., Vijayakumar, S., & Wolpert, D. M. (2014). On the origins of suboptimality in human probabilistic inference. *PLoS Computational Biology*, 10(6), e1003661.
- Acuna, D. E., Berniker, M., Fernandes, H. L., & Kording, K. P. (2015). Using psychophysics to ask if the brain samples or maximizes. *Journal of Vision*, 15(3), 7. <https://doi.org/10.1167/15.3.7>.
- Anderson, N. H. (1967). Averaging model analysis of set-size effect in impression formation. *Journal of Experimental Psychology*, 75(2), 158-165.

## Author note

This work was supported by an Economic and Social Research Council grant (ES/K004948/1) and a European Research Council consolidator grant (817492-SAMPLING) to ANS and an Economic and Social Research Council grants (ES/K002201/1 and ES/N018192/1) and a Leverhulme Trust grant (RP2012-V-022) grant to NS. The authors thank Jerome Busemeyer and Richard Shiffrin for helpful discussions. The data as well as analysis code from all experiments is available on the Open Science Framework: <http://doi.org/10.17605/OSF.IO/8QS6J>.

- Anderson, N. H. (1981). *Foundations of information integration theory*. New York: Academic Press.
- Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., & Quattrocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3), 554–559. <https://doi.org/10.1073/pnas.1517441113>.
- Denrell, J., Sanborn, A. N., & Spicer, J. (2019). Implicit corrections for missing feedback: imputation vs. statistical models.
- Drugowitsch, J., Wyart, V., Devauchelle, A.-D., & Koehlin, E. (2016). Computational precision of mental inference as critical source of human choice suboptimality. *Neuron*, 92(6), 1–14. <https://doi.org/10.1016/j.neuron.2016.11.005>.
- Duval, S., & Tweedie, R. (2000). A nonparametric “trim and fill” method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, 95(449), 89–98.
- Elwin, E., Juslin, P., Olsson, H., & Enkvist, T. (2007). Constructivist coding learning from selective feedback. *Psychological Science*, 18(2), 105–110.
- Fein, S., & Hilton, J. L. (1992). Attitudes toward groups and behavioral intentions toward individual group members: The impact of nondiagnostic information. *Journal of Experimental Social Psychology*, 28(2), 101–124.
- Fernbach, P. M., Darlow, A., & Sloman, S. A. (2010). Neglect of alternative causes in predictive but not diagnostic reasoning. *Psychological Science*, 21(3), 329–336.
- Fernbach, P. M., Darlow, A., & Sloman, S. A. (2011). Asymmetries in predictive and diagnostic reasoning. *Journal of Experimental Psychology: General*, 140(2), 168–185.
- Gershman, S., & Goodman, N. (2014). Amortized inference in probabilistic reasoning. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 36).
- Glover, S. M. (1997). The influence of time pressure and accountability on auditors' processing of nondiagnostic information. *Journal of Accounting Research*, 35(2), 213–226.
- Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, 21, 263–268.
- Hackenbrack, K. (1992). Implications of seemingly irrelevant evidence in audit judgment. *Journal of Accounting Research*, 126–136.
- Henriksson, M. P., Elwin, E., & Juslin, P. (2010). What is coded into memory in the absence of outcome feedback? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(1), 1–16.
- Hilton, J. L., & Fein, S. (1989). The role of typical diagnosticity in stereotype-based judgments. *Journal of Personality and Social Psychology*, 57(2), 201–211.
- Hotelling, J. M., Cohen, A. L., Shiffrin, R. M., & Bussemeyer, J. R. (2015). The dilution effect and information integration in perceptual decision making. *PLoS One*, 10(9), e0138481.
- Igou, E. R. (2007). Additional thoughts on conversational and motivational sources of the dilution effect. *Journal of Language and Social Psychology*, 26(1), 61–68.
- Igou, E. R., & Bless, H. (2005). The conversational basis for the dilution effect. *Journal of Language and Social Psychology*, 24(1), 25–35.
- Johnson, S. G. B., Rajeev-Kumar, G., & Keil, F. C. (2016). Sense-making under ignorance. *Cognitive Psychology*, 89, 39–70. <https://doi.org/10.1016/j.cogpsych.2016.06.004>.
- Johnston, A. M., Johnson, S. G. B., Koven, M. L., & Keil, F. C. (2017). Little Bayesians or little Einsteins? Probability and explanatory virtue in children's inferences. *Developmental Science*, 20(6), e12483.
- Juslin, P., Nilsson, H., & Winman, A. (2009). Probability theory, not the very guide of life. *Psychological Review*, 116(4), 856–874.
- Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3, 430–454.
- Kemmelmeier, M. (2004). Separating the wheat from the chaff: Does discriminating between diagnostic and nondiagnostic information eliminate the dilution effect? *Journal of Behavioral Decision Making*, 17(3), 231–243.
- Kemmelmeier, M. (2007a). Does the dilution effect have a conversational basis? *Journal of Language and Social Psychology*, 26(1), 48–60.
- Kemmelmeier, M. (2007b). Is diagnostic evidence on the dilution effect weakened when nondiagnostic objections are added? A response to Igou (2007). *Journal of Language and Social Psychology*, 26(1), 69–74.
- Khmelani, S. S., Sussman, A. B., & Oppenheimer, D. M. (2011). Harry potter and the sorcerer's scope: latent scope biases in explanatory reasoning. *Memory & Cognition*, 39(3), 527–535.
- Labella, C., & Koehler, D. J. (2004). Dilution and confirmation of probability judgments based on nondiagnostic evidence. *Memory & Cognition*, 32(7), 1076–1089.
- Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data*. John Wiley & Sons.
- Meder, B., & Mayrhofer, R. (2017). Diagnostic causal reasoning with verbal information. *Cognitive Psychology*, 96, 54–84.
- Meder, B., Mayrhofer, R., & Waldmann, M. R. (2014). Structure induction in diagnostic causal reasoning. *Psychological Review*, 121(3), 277–301.
- Meyvis, T., & Janiszewski, C. (2002). Consumers' beliefs about product benefits: The effect of obviously irrelevant product information. *Journal of Consumer Research*, 28(4), 618–635.
- Murphy, G. L., & Ross, B. H. (1994). Predictions from uncertain categorizations. *Cognitive Psychology*, 27, 148–193.
- Murphy, G. L., & Ross, B. H. (2010). Uncertainty in category-based induction: When do people integrate across categories? *Journal of Experimental Psychology: Learning Memory, and Cognition*, 36(2), 263–276.
- Nisbett, R. E., Zukier, H., & Lemley, R. E. (1981). The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. *Cognitive Psychology*, 13(2), 248–277.
- Nosofsky, R. M. (1986). Attention, similarity and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115, 39–57.
- Nosofsky, R. M. (1992). *Exemplars, prototypes, and similarity rules. From learning theory to connectionist theory: Essays in honor of William K. Estes*. Hillsdale, New Jersey: Lawrence Erlbaum Associates 149–167.
- Nosofsky, R. M. (2015). An exemplar-model account of feature inference from uncertain categorizations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41(6), 1929–1941.
- Peters, E., & Rothbart, M. (2000). Typicality can create, eliminate, and reverse the dilution effect. *Personality and Social Psychology Bulletin*, 26(2), 177–187.
- Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, 72(3), 346–354.
- Ramachandran, V. S. (1992). Filling in gaps in perception: Part I. *Current Directions in Psychological Science*, 1(6), 199–205.
- Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16, 225–237.
- Sanborn, A. N., & Beierholm, U. R. (2016). Fast and accurate learning when making discrete numerical estimates. *PLoS Computational Biology*, 12(4), 1–28. <https://doi.org/10.1371/journal.pcbi.1004859>.
- Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, 20(12), 883–893.
- Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to the rational model of categorization. *Psychological Review*, 117, 1144–1167.
- Shanteau, J. (1975). Averaging versus multiplying combination rules of inference judgment. *Acta Psychologica*, 39(1), 83–89.
- Shelton, S. W. (1999). The effect of experience on the use of irrelevant evidence in auditor judgment. *The Accounting Review*, 74(2), 217–224.
- Tetlock, P. E., & Boettger, R. (1989). Accountability: a social magnifier of the dilution effect. *Journal of Personality and Social Psychology*, 57(3), 388–398.
- Tetlock, P. E., Lerner, J. S., & Boettger, R. (1996). The dilution effect: judgmental bias, conversational convention, or a bit of both? *European Journal of Social Psychology*, 26, 915–934.
- Troutman, C. M., & Shanteau, J. (1977). Inferences based on nondiagnostic information. *Organizational Behavior and Human Performance*, 19(1), 43–55.
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84, 327–352.
- Tversky, A., & Kahneman, D. (1983). Extensional vs intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90, 293–315.
- Waller, W. S., & Zimelman, M. F. (2003). A cognitive footprint in archival data: Generalizing the dilution effect from laboratory to field settings. *Organizational Behavior and Human Decision Processes*, 91(2), 254–268.
- Wallsten, T. S. (1976). A note on Shanteau's 'Averaging versus multiplying combination rules of inference judgment'. *Acta Psychologica*, 40(4), 325–330.
- Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science*, 6, 1–15.
- Yurovsky, D., Boyer, T. W., Smith, L. B., & Yu, C. (2013). Probabilistic cue combination: less is more. *Developmental Science*, 16(2), 149–158.
- Zukier, H. (1982). The dilution effect: The role of the correlation and the dispersion of predictor variables in the use of nondiagnostic information. *Journal of Personality and Social Psychology*, 43(6), 1163–1174.
- Zukier, H., & Jennings, D. L. (1984). Nondiagnosticity and typicality effects in prediction. *Social Cognition*, 2(3), 187–198.