Check for updates

Research Article

Psychological Science
1–13
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0956797618792256
www.psychologicalscience.org/PS

The Binary Bias: A Systematic Distortion in the Integration of Information

Matthew Fisher[1] and Frank C. Keil[2]
[1]Department of Social and Decision Sciences, Carnegie Mellon University, and
[2]Department of Psychology, Yale University

Abstract

One of the mind's most fundamental tasks is interpreting incoming data and weighing the value of new evidence. Across a wide variety of contexts, we show that when summarizing evidence, people exhibit a binary bias: a tendency to impose categorical distinctions on continuous data. Evidence is compressed into discrete bins, and the difference between categories forms the summary judgment. The binary bias distorts belief formation—such that when people aggregate conflicting scientific reports, they attend to valence and inaccurately weight the extremity of the evidence. The same effect occurs when people interpret popular forms of data visualization, and it cannot be explained by other statistical features of the stimuli. This effect is not confined to explicit statistical estimates; it also influences how people use data to make health, financial, and public-policy decisions. These studies (N = 1,851) support a new framework for understanding information integration across a wide variety of contexts.

Keywords

biases and heuristics, judgment and decision making, categorical thinking, information integration, open data, open materials

Received 5/28/18; Revision accepted 6/9/18

We often confront evidence and must construct a summary representation of the data. For example, some sources claim caffeine has adverse health effects, while others claim it has health benefits. These multiple reports, each with varying degrees of extremity in their claims, must be combined and then used to inform decisions. Judgments are similarly formed about political topics, attitudes toward other people, and many other cases that require integration of multiple values along an implicit or explicit scale. Here, we demonstrate that across a wide variety of contexts, people show a persistent bias to dichotomize evidence.

To illustrate the phenomenon visually, consider the graphical depiction in Figure 1. Given the values in the graphs, people estimate that the restaurant whose customer distribution is shown on the right has nearly one person more per table than the restaurant whose customer distribution is shown on the left. In fact, both restaurants had the exact same true mean of three people per table. We propose that a systematic evidence-weighting error, the *binary bias*, helps explain why people incorrectly estimate these averages. Through the following series of studies, we demonstrate that the binary bias not only affects how consumers interpret online ratings (Fisher, Newman, & Dhar, 2018) but also helps characterize how evidence is construed more generally across many different settings and tasks.

We propose that people neglect the relative strength of evidence and instead treat evidence as binary. In the case of the graphs in Figure 1, people intuitively compress continuous data into a "binary" format, estimating the mean on the basis of whether there are more data on the left-hand or the right-hand side of each graph, without taking into account that the bars closer to the midpoint (the "2" and "4" bars) are less extreme than the bars farther from the midpoint (the "1" and "5" bars).

In the current studies, binary thinking was operationalized through a statistic dubbed an *imbalance score*. The imbalance score is the difference in total data points on one side of the boundary (e.g., the "3" bar) versus the other. A bottom-heavy distribution such

Corresponding Author:
Matthew Fisher, Carnegie Mellon University, Department of Social and Decision Sciences, 5000 Forbes Ave., Pittsburgh, PA 15213
E-mail: mcfisher@cmu.edu

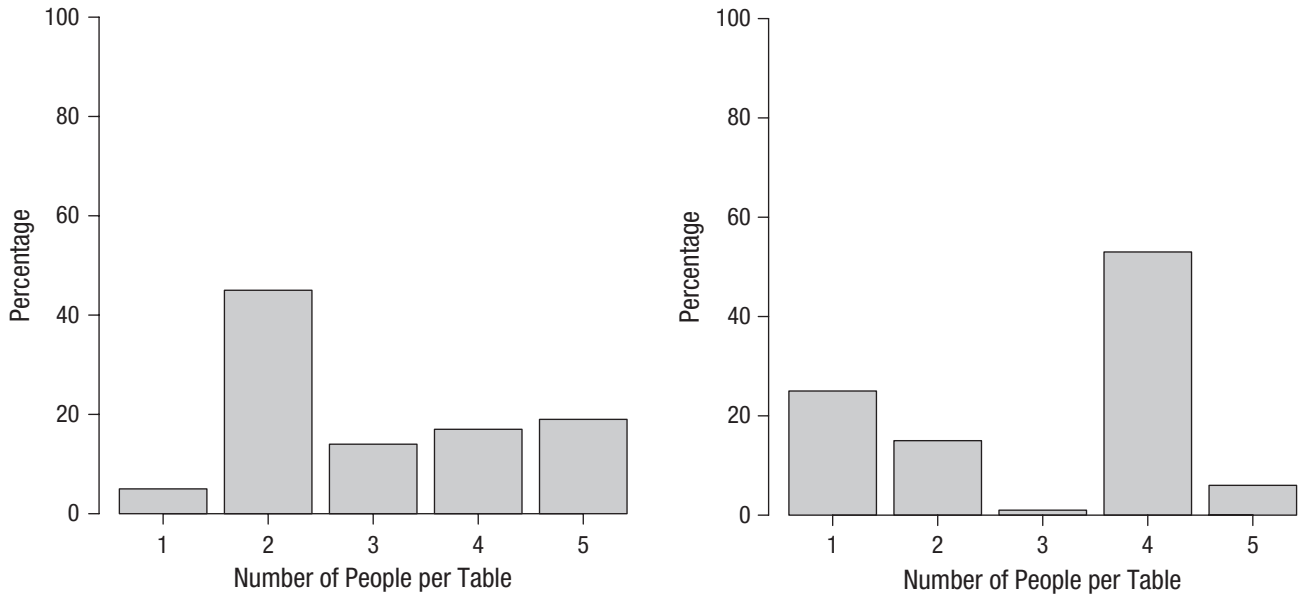


Fig. 1. An illustration of the binary bias. For both distributions, the mean is 3.00, but people estimate the distribution mean on the left as 2.64 and on the right as 3.46.

as the one depicted in the left-hand graph in Figure 1 has a highly negative imbalance score because the sum of the "1" bar and "2" bar is greater than the sum of the "4" bar and "5" bar. A top-heavy distribution such as the one depicted in the right-hand graph has a highly positive imbalance score because the sum of the "4" bar and "5" bar is greater than the sum of the "1" bar and "2" bar. The imbalance score treats the continuous range of values as binary and compares the difference between the two categories. Thus, this statistic can be used to assess whether binary thinking underlies people's intuitive summaries.

This account shares similarities with another proposed pattern of integration: the *tallying heuristic*, which involves summing the total number of cues that favor one option over another across multiple dimensions (Gigerenzer & Goldstein, 1996). Unlike the tallying heuristic, the binary bias posits categorical thinking along a single dimension and that people discretize a continuous value range.

Information Integration

Intuitive descriptive and inferential statistics systematically deviate from the normative models of probability (Peterson & Beach, 1967; but see Griffiths & Tenenbaum, 2006), yet estimates of central tendency have been found to be quite accurate (Beach & Swenson, 1966; Spencer, 1961). Our proposal offers an alternative account: Namely, when distributions are imbalanced, people's estimates of the mean will be biased. But what are the cognitive processes involved in generating these summaries? Researchers have explored the "cognitive algebra" underlying information integration, such as additive (Betsch, Plessner, Schwieren, & Gütig, 2001) or weighted-averaging accounts (Anderson, 1981; Manis, Gleason, & Dawes, 1966; Rosenbaum & Levin, 1968; see also Betsch, Kaufmann, Lindow, Plessner, & Hoffmann, 2006). Here, we posit a new model: Unlike these algebraic models of integration, our account is based on categorical thinking.

The binary bias not only applies to explicitly statistical presentations, as in Figure 1, but also contributes to our understanding of how information is acquired across time. Previous work has suggested that summaries over time provide further support for the averaging model (Kahneman, Fredrickson, Schreiber, & Redelmeier, 1993). A variant of the averaging principle, the *peak-end rule*, posits that people evaluate experiences by averaging the height (peak) and finish (end) of an episode. Similar patterns occur when people recall serially presented information: Memories for the first and last items are most accurate (Henson, 1998; Peters & Bijmolt, 1997). These order effects can be explained by salience, which also leads high magnitude values to be disproportionately weighted (Tsetsos, Chater, & Usher, 2012). In these ways, the process of integrating sequentially presented information is subject to distortions. These previous findings provide benchmarks to test against the binary bias.

Binary Thinking

Our proposal is that binary thinking constrains the process of information integration. Sorting continuous data into separate categories reduces complexity and allows for efficient information processing (Smith & Medin, 1981). The effects of categorical thinking have been found across many areas of psychology.

In perception, the visual system selects a particular interpretation of ambiguous stimuli and neglects other possibilities, even when it leads to suboptimal decision making (Fleming, Maloney, & Daw, 2013; Harnad, 1987). Similar effects arise across high-level cognition. People treat continuous evidence dichotomously rather than representing degrees of belief in domains such as categorization (Murphy & Ross, 1994), causal reasoning (Johnson, Merchant, & Keil, 2015; Soo & Rottman, 2018), stereotyping (Corneille & Judd, 1999), and economic decision making (Isaac & Schindler, 2013). People overly focus on a single causal mechanism when estimating the probability of an effect (Fernbach, Darlow, & Sloman, 2011), contrary to many Bayesian models of cognition. Similarly, people systematically overweight the strength of a particular piece of evidence (e.g., size of the effect) and neglect its weight (e.g., sample size; Griffin & Tversky, 1992; Kvam & Pleskac, 2016). Categorical thinking also leads trained researchers to make false dichotomizations around "$p = .05$" (McShane & Gal, 2015).

The Current Studies

Our core claim is that categorical thinking distorts information integration. We propose that when considering evidence, including graphical depictions, people display a binary bias: They treat evidence as all or none without tracking the differential impact of graded evidence. Given that many alternative accounts related to the binary bias have been addressed with process evidence (Fisher et al., 2018), the goal of the present research was to demonstrate the scope of the bias and to propose it as a general framework for understanding evidence aggregation.

Study 1a

Attitudes are commonly understood as evaluative summaries of available evidence (Banaji & Heiphetz, 2010). People must make a summary judgment of the conflicting information they encounter. In Study 1, we tested whether attitude formation reflects a binary bias.

Method

Participants. Four separate groups of participants were recruited for Study 1a. The four groups consisted of 154 participants (82 male; age: $M = 36.54$ years, $SD = 12.54$), 152 participants (71 male; age: $M = 34.67$ years, $SD = 11.69$), 152 participants (70 male; age: $M = 34.80$ years, $SD = 11.36$), and 147 participants (65 male; age: $M = 36.41$ years, $SD = 12.28$). All were from the United States, and all completed the study through Amazon Mechanical Turk. A power analysis with an effect size (f^2) of .09 based on pilot testing estimated that 150 participants would be needed in each group to detect an effect (power = .95). Participants did not complete multiple studies; each study contained a unique, naive sample, and once the requested number of participants completed each study, data collection ended. Informed consent was obtained from all participants in all studies.

Materials and procedure. Participants in Study 1 were divided into four groups; each group considered a randomly assigned topic from one of four domains: scientific reports, eyewitness testimonies, social judgments, or consumer reviews (see Appendix S1 in the Supplemental Material available online for the full sets of topics for each of the four domains). For example, a participant considering scientific reports would see statements about whether or not a new medication leads to feelings of hunger. Participants viewed a series of claims about the relationship between the two variables. There were five levels of evidence that participants could see for each topic: strong positive evidence (e.g., "One group of scientists found that the new medication makes feeling hungry *4 times more* likely"), weak positive evidence (e.g., "One group of scientists found that the new medication makes feeling hungry *2 times more* likely"), neutral evidence (e.g., "One group of scientists found that the new medication *does not change the likelihood* of feeling hungry"), weak negative evidence (e.g., "One group of scientists found that the new medication makes feeling hungry *2 times less* likely"), and strong negative evidence (e.g., "One group of scientists found that the new medication makes feeling hungry *4 times less* likely"). Each participant viewed a sequence of 17 total instances of the five levels of evidence. One claim would appear on the screen (e.g., "One group of scientists found that the new medication makes feeling hungry *4 times more* likely"), and participants could not click to continue until 5 s had elapsed. Next, another claim about the same topic would appear, and this process continued until participants had viewed all 17 items. The same level of evidence (i.e., the same statement) could appear multiple times within a given sequence.

As in the example above, evidence in the medical domain was presented in terms of likelihoods, but for generalizability, the levels of evidence were formatted differently in each of the other domains. For example,

in the eyewitness testimony domain, evidence was presented as confidence percentages ("One witness is 100% confident that the defendant did not commit crime X," "One witness is 50% confident that the defendant did not commit crime X," etc.). See Appendix S2 in the Supplemental Material for full details of the levels of evidence used across all four domains.

The distributions of levels of evidence (total number of strong positive, weak positive, neutral, weak negative, and strong negative) were constructed such that the full set of stimuli covered the widest possible range of imbalance scores (−5 to 5). As mentioned in the introduction, the imbalance score is computed by subtracting the amount of evidence below the midpoint from the amount of evidence above the midpoint. For the sequences in Study 1, the imbalance scores equaled the number of strong and weak negative evidence scores subtracted from the number of strong and weak positive evidence scores. Distributions of evidence were randomly generated and then selected so that there were three distributions for each imbalance score from −5 to 5, for a total of 33 distributions. See Appendix S3 in the Supplemental Material for the set of distributions used as the sequences of evidence in Study 1. The true weighted average of all distributions totaled to no effect. Each participant viewed one sequence of 17 statements.

After viewing all 17 pieces of evidence for their randomly assigned topic, participants were asked to summarize the evidence they had just seen. For example, after viewing 17 claims about the relationship between a new medication and feeling hungry, they were asked, "How much does the new medication change the likelihood of feeling hungry?" Participants responded on 9-point Likert scales (e.g., from "4 times less likely" to "4 times more likely"). The anchors on the Likert scale changed on the basis of the wording of the evidence, as listed in Appendix S2.

Results

The effect of imbalance did not change across domains, so results from all four contexts are reported together. A linear regression model predicted participants' summary judgments using imbalance score, mode (a measure of the most salient level of evidence), first level of evidence presented (primacy), and last level of evidence presented (recency). While controlling for these other variables, imbalance score predicted participant summary judgments, $\beta = 0.31$, $SE = 0.04$, $b = 4.62$, $SE = 0.63$, 95% confidence interval (CI) = [3.38, 5.85], $p < .001$. Additionally, we found that the first piece of evidence viewed by participants was also a significant predictor, $\beta = 0.08$, $SE = 0.04$, $b = 2.70$, $SE = 1.26$, 95% CI = [0.23, 5.17], $p = .03$. The most frequently appearing level of evidence, $\beta = 0.06$, $SE = 0.04$, $b = 2.40$, $SE = 1.62$, 95% CI = [−0.77, 5.58], $p = .14$, and the last piece of evidence, $\beta = 0.01$, $SE = 0.04$, $b = 0.40$, $SE = 1.29$, 95% CI = [−2.13, 2.94], $p = .75$, were not predictive of participants' summary judgments. These results demonstrate across a wide array of domains, using different forms of evidence, that the binary bias has a stronger influence on the formation of beliefs and attitudes than the previously documented factors of order and salience.

Study 1b

We next arranged the presentation order to elicit stronger order effects. This provided a more stringent test for assessing the distinct influence of imbalance.

Method

Participants. One hundred forty-nine participants (81 male; age: $M = 35.81$ years, $SD = 10.70$) from the United States completed the study through Amazon Mechanical Turk.

Materials and procedure. Since domain did not interact with the effect of imbalance score in Study 1a, we randomly selected one domain (social) to be used in Study 1b. The procedure for Study 1b was identical to the procedure for Study 1a, except that instead of all 17 statements appearing in a random order, each level of evidence (strong negative, weak negative, no effect, weak positive, strong positive) was grouped together, and then those five blocks were presented in a random order. Grouping the statements by evidence type was designed to accentuate primacy and recency effects. After viewing all 17 statements, participants estimated the relationship between the two variables on a 9-point Likert scare (from *extremely unlikely* to *extremely likely*).

Results

After analyses emphasized the order of evidence, a linear regression model found that imbalance scores remained a marginally significant predictor of participants' summary judgments, $\beta = 0.16$, $SE = 0.09$, $b = 2.16$, $SE = 1.18$, 95% CI = [−0.16, 4.49], $p = .07$. There was no effect of the most frequently appearing level of evidence, $\beta = 0.004$, $SE = 0.09$, $b = 0.13$, $SE = 2.99$, 95% CI = [−5.77, 6.04], $p = .96$; the first-viewed level of evidence, $\beta = -0.12$, $SE = 0.09$, $b = -3.35$, $SE = 2.46$, 95% CI = [−8.23, 1.52], $p = .18$; or the last-viewed level of evidence, $\beta = -0.06$, $SE = 0.08$, $b = -1.80$, $SE = 2.42$, 95% CI = [−6.58, 2.99], $p = .46$. This pattern of results

indicated that even when order is made more salient by grouping levels of evidence, binary thinking still influences summary judgments.

Study 2a

Study 1 showed that the binary bias influenced summaries of beliefs formed over time; however, decisions based on experience can be made differently than decisions based on descriptions (Hertwig, Barron, Weber, & Erev, 2004). To explore this difference and test the generalizability of the binary bias, we next tested how people interpret data presented in graphical form. If the binary bias applies to information integration more generally, then imbalance scores should predict participants' intuitive estimates of a distribution's mean.

Method

Participants. Two hundred thirty-eight participants (129 male; age: M = 35.38 years, SD = 11.55) from the United States completed the study through Amazon Mechanical Turk. On the basis of pilot testing, a power analysis determined that approximately 20 participants rating each item would be required to detect a small-sized effect (power = .95).

Materials and procedure. Across science, business, and popular media, data visualization is ubiquitous, yet relatively little is known about how data are intuitively understood (Spiegelhalter, Pearson, & Short, 2011), and even people with statistical training make systematic errors (Ibrekk & Morgan, 1987). In Study 2a, we used vertical bar charts to test the hypothesis that the binary bias, as operationalized by the imbalance score, can predict estimates of the mean. The stimuli in Study 2a consisted of 120 histograms, each depicting five bars labeled "1" to "5" (see Fig. 1). The stimuli were divided into three sets: 40 of the graphs had a mean of 2.75, 40 had a mean of 3.00, and 40 had a mean of 3.25. For each set, the 40 distributions were randomly and uniformly selected from all possible distributions with that set's mean. Each distribution's five bins totaled 100 and had at least one data point in each bin.

Participants were divided into three groups, and each group rated one of the three sets of graphs. Each participant evaluated a random subset of 10 of the 40 graphs in his or her assigned set; thus, each graph was rated approximately 20 times. Participants were asked to estimate the average of each distribution they were presented. Simply asking the "average" is ambiguous, because one can estimate the mean y-value or the mean x-value of a vertical bar chart. To encourage the appropriate interpretation, we asked participants, "Based on your immediate judgment, on average, how many people sat per table?" Participants responded on a sliding scale from 1 to 5, labeled every 0.5 points. The slider's current value was displayed to the hundredths decimal place next to the sliding scale. Furthermore, the x-axis of each graph was labeled "Number of People Per Table" and the y-axis was labeled "Percentage (%)." The context, as well as the sliding scale's range of values, made it clear to participants that they should estimate the mean x-value.

Although in statistics the term "average" captures multiple measures of central tendency, we used the word "average" in the dependent measure because it is colloquially understood as the arithmetic mean (Goldstein & Rothschild, 2014). Participants were asked to use their "immediate judgment" in order to elicit intuitive responses and discourage them from taking the time to mathematically compute the average using the numbers depicted on the graph.

Results

Although all 40 graphs in each set had an identical mean, the total of the lower value bars ("1"s and "2"s) and the higher value bars ("4"s and "5"s) differed across graphs. If participants treated the evidence in the graphs as binary, then they should have based their estimations on the degree to which one side of the graph contained higher bars, regardless of the relative strength of the data (e.g., weighting the "4" bar as much as the "5" bar). To assess whether participants' estimates were driven by binary thinking, we again computed an imbalance score for each distribution. The effect of imbalance did not vary as a function of the distributions' true mean, so results from the three sets of stimuli are reported together.

As previously mentioned, well-studied factors such as salience (Tsetsos et al., 2012) and the median (Peterson & Miller, 1964) may also influence how people summarize evidence. Thus, in Study 2a, salience (the tallest bar) and the median were included as independent factors in our analysis and used as benchmarks against which the strength of the binary bias could be measured. Additionally, we included the "peakedness" (kurtosis), spread (SD), and arithmetic mean in the model. Note that skewness was omitted from this and subsequent models because of its correlation with imbalance (r = −.83; variance inflation factor, or VIF > 4). Previous research has provided evidence against skewness as an explanation for the binary bias (Fisher et al., 2018).

A linear mixed-effects regression model from the lme4 package in the R programming environment

Table 1. Mixed-Effects Regression Results for Mean Estimates in Study 2a

Fixed effect	*b*	*SE*	β	*SE* β	Bootstrapped 95% CI for *b*
Intercept	0.96	0.11	−0.01	0.03	[0.05, 1.76]
Imbalance	0.01	0.001	0.20***	0.04	[0.004, 0.01]
Mean	0.57	0.11	0.19***	0.04	[0.31, 0.80]
Mode	0.05	0.01	0.10***	0.02	[0.03, 0.07]
Kurtosis	0.07	0.03	0.08**	0.03	[0.03, 0.12]
Standard deviation	0.12	0.07	0.05	0.03	[−0.01, 0.26]
Median	−0.04	0.03	−0.04	0.03	[−0.10, 0.01]

Note: For these models, the number of observations was 2,368, the number of subjects was 238, and the number of items was 120. CI = confidence interval.
**$p < .01$. ***$p < .001$.

(Bates, Maechler, Bolker, & Walker, 2015) predicted participants' mean estimates, with imbalance score and mode as fixed effects. The model included random intercepts for subjects and random slopes for by-subject differences in the effect of imbalance scores on mean estimates. In addition to reporting *p* values, we tested whether bootstrapped 95% CIs for coefficients included zero (Bates, Mächler, Bolker, & Walker, 2014). We found that, controlling for other statistical features of the distribution, the imbalance score was a strong predictor of participants' estimates, β = 0.20, *SE* = 0.04, *b* = 0.01, *SE* = 0.01, bootstrapped 95% CI = [0.004, 0.01], $p < .001$. Independently, the mean, β = 0.19, *SE* = 0.04, *b* = 0.57, *SE* = 0.11, bootstrapped 95% CI = [0.31, 0.80], $p < .001$; the mode (tallest bar), β = 0.10, *SE* = 0.02, *b* = 0.05, *SE* = 0.01, bootstrapped 95% CI = [0.03, 0.07], $p < .001$; and kurtosis (peakedness), β = 0.08, *SE* = 0.03, *b* = 0.07, *SE* = 0.03, bootstrapped 95% CI = [0.03, 0.12], $p = .004$, also predicted participants' ratings, suggesting multiple inputs to the mean estimation process. See Table 1 for details of the regression model. These results demonstrate that, above and beyond other heuristics such as a reliance of salience, the binary bias helps explain how people aggregate data presented as a histogram.

Study 2b

If participants' estimates were perfectly accurate in Study 2a, then they would be identical for every item. This could create experimental demand to vary responses, leading a weak tendency toward binary thinking to be exaggerated. Thus, in Study 2b, participants viewed distributions with different true means. To ensure participants were as accurate as possible, we financially incentivized performance.

Method

Participants. Two hundred twenty-six participants (104 male; age: *M* = 37.28 years, *SD* = 11.77) from the United States completed the study through Amazon Mechanical Turk.

Materials and procedure. In Study 2b, we used the same procedure as in Study 2a, with a few alterations. First, instead of separate groups of participants rating each set of graphs, each participant estimated the mean for a random subset of 30 graphs drawn from all three sets of stimuli. Thus, the true mean of the graphs changed from trial to trial. Second, to motivate participants to be as accurate as possible, they were informed that they would be eligible for a bonus payment if they outperformed other participants. To avoid participants' calculating the mean in order to earn the additional payment, we displayed the graph for only 5 s before participants estimated the mean on a separate page. Even with the time limit and auto-advance feature, participants completed most trials (*M* = 29.06 out of 30 trials).

Results

Replicating the results from Study 2a, the results of Study 2b showed that participants' estimates were again predicted by the imbalance score (see Table 2 and Fig. 2), suggesting that the binary bias is an important component of intuitively extracting summary statistics. These factors remained strong predictors even after we controlled for other statistical features of the graphs. As in Study 2a, mean and mode were independent predictors of participants' responses. See Table 2 for details of the regression model. Importantly, the effect found in Study 2a was replicated even when experimental demand was removed and participants were motivated to give accurate responses.

Study 3a

We next explored whether the binary bias requires certain graphical features to be present and, furthermore, the extent to which the binary bias is a visual

Table 2. Mixed-Effects Regression Results for Mean Estimates in Study 2b

Fixed effect	*b*	*SE b*	β	*SE* β	Bootstrapped 95% CI for *b*
Intercept	1.40	0.21	0.00	0.03	[1.02, 1.83]
Imbalance	0.01	0.001	0.14***	0.03	[0.004, 0.01]
Mean	0.39	0.06	0.12***	0.02	[0.27, 0.51]
Mode	0.13	0.01	0.22***	0.01	[0.11, 0.14]
Kurtosis	0.02	0.02	0.02	0.02	[–0.02, 0.05]
Standard deviation	0.10	0.05	0.04*	0.02	[0.01, 0.20]
Median	–0.02	0.02	–0.02	0.02	[–0.06, 0.01]

Note: For these models, the number of observations was 6,567, the number of subjects was 226, and the number of items was 120. CI = confidence interval.
*$p < .05$. ***$p < .001$.

versus a cognitive illusion. In Study 3a, participants estimated the average of a series of distributions presented in a variety of formats.

Method

Participants. Three hundred twenty-one participants (172 male; age: $M = 35.42$ years, $SD = 11.78$) from the United States completed the study through Amazon Mechanical Turk.

Materials and procedure. In Study 3a, each participant viewed data in one of four formats (see Fig. 3). First, we presented the data in horizontal bar charts. If the effect arises from an intuitive sense of physical balance (Siegler, 1976), as present in the vertical bar charts from Study 2, then we would no longer expect to see a bias when participants considered the horizontal bar chart. Second, we presented the data in pie charts, which are interpreted differently than bar charts (Simkin & Hastie, 1987). If the effect arises from the spatial relations between the representations of each value, then we would no longer expect to see a bias for pie charts because the position of each slice in a pie chart is arbitrary and nonmonotonic. Third, we tested whether the effect would arise even in the absence of graphical representation by presenting the data as verbal descriptions. No visual or spatial cues could trigger the bias in this context. Lastly, since the verbal description used percentage signs, which have been shown to be especially difficult for people to process (Gigerenzer & Hoffrage, 1995), we also presented the data as verbal descriptions without percentage signs. If the binary bias persisted even for verbal descriptions, it would suggest that the effect is not only a visual illusion but also a cognitive bias.

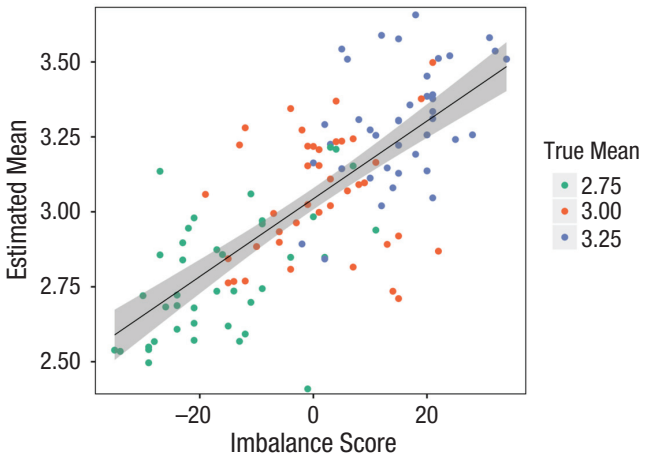
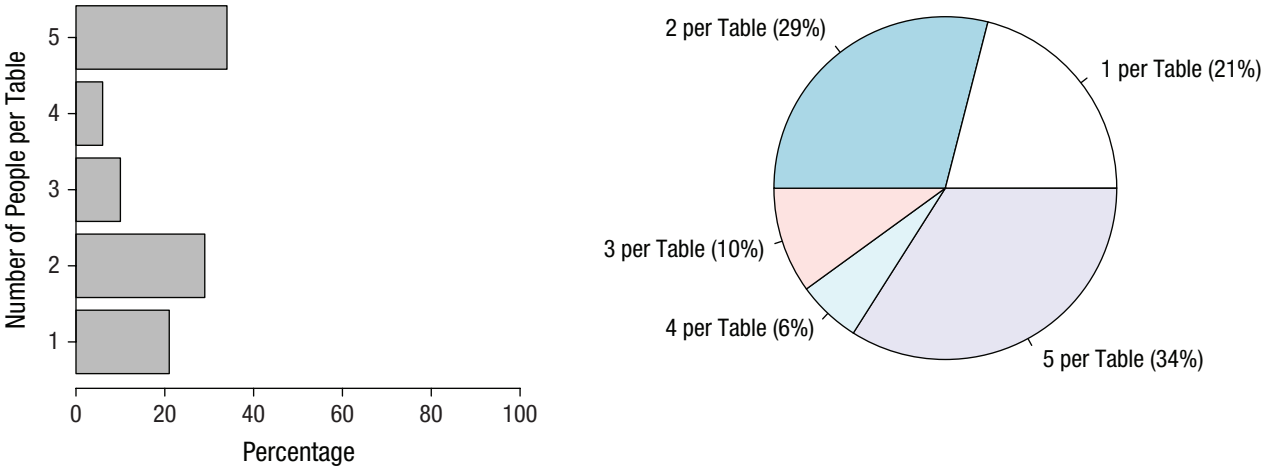


Fig. 2. Scatterplot (with best-fitting regression line) showing the relationship in Study 2b between imbalance score and participants' estimates of the mean, separately for each of the three true mean values. The error band shows the 95% confidence interval.

To test whether the errors found in Study 2 generalized to other formats of presentation, we first randomly selected one of the three sets of stimuli from Study 2, and the set of 40 distributions with a mean of 3.25 was chosen. These 40 distributions were depicted in four new formats. Participants were divided into four groups, and each group was presented with distributions in one of the four formats. Each participant viewed a random subset of 10 of the 40 items in his or her assigned format. In Study 3a, we used the same procedure, including the framing and dependent measure, from Study 2a.

Results

Estimates of the mean for the same distributions presented in different formats were strongly correlated, indicating that the errors made by participants were systematic and not due to random noise (see Table 3). Across formats, the binary bias again explained estimates of the mean above and beyond the influence of other statistical features. A linear mixed-effects model predicted participants' estimates with imbalance score, mode, standard deviation, median, and format as fixed



34% had 5 per Table
6% had 4 per Table
10% had 3 per Table
29% had 2 per Table
21% had 1 per Table

34 had 5 per Table
6 had 4 per Table
10 had 3 per Table
29 had 2 per Table
21 had 1 per Table

Fig. 3. The two visual formats (top) and two verbal formats (bottom) of presentation in Study 3a.

effects and included random intercepts for subject and by-subject random slopes for imbalance. Kurtosis was not included because of its strong correlation with standard deviation ($r = -.82$, VIF > 4). Participants' estimates were significantly predicted by the imbalance score, $\beta = 0.18$, $SE = 0.02$, $b = 0.01$, $SE = 0.001$, bootstrapped 95% CI = [0.009, 0.01], $p < .001$. The mode, $\beta = 0.11$, $SE = 0.02$, $b = 0.07$, $SE = 0.01$, bootstrapped 95% CI = [0.05, 0.09], $p < .001$, and standard deviation, $\beta = 0.09$, $SE = 0.02$, $b = 0.21$, $SE = 0.05$, bootstrapped 95% CI = [0.11, 0.30], $p < .001$, were also significant predictors. Using a likelihood-ratio test, we compared the model's goodness of fit with that of a second identical model, which also included the Imbalance Score × Stimuli Format interaction term as a fixed effect. This test revealed no significant difference between the models, $\chi^2(4) = 4.80$, $p = .31$, suggesting that the effect of imbalance did not depend on the format.

Study 3b

Unlike the previous formats, in a dot plot, all the raw data are visible and not neatly grouped into discrete bins. In Study 3b, we tested whether the binary bias would persist without clear categories being present.

Method

Participants. Seventy-nine participants (45 male; age: $M = 35.15$ years, $SD = 12.12$) from the United States completed the study through Amazon Mechanical Turk.

Table 3. Correlations Between Stimuli Formats in Study 3a

Format	Vertical bar chart	Horizontal bar chart	Pie chart	Verbal (with %)
Vertical bar chart	—			
Horizontal bar chart	.63*** [.40, .63]	—		
Pie chart	.43* [.14, .65]	.73*** [.54, .85]	—	
Verbal (with %)	.46**[.17, .68]	.63** [.40, .79]	.66*** [.44, .80]	—
Verbal (without %)	.38* [.08, .62]	.46** [.18, .68]	.54*** [.28, .73]	.38* [.08, .62]

Note: Values in brackets are 95% confidence intervals.
*$p < .05$. **$p < .01$. ***$p < .001$.

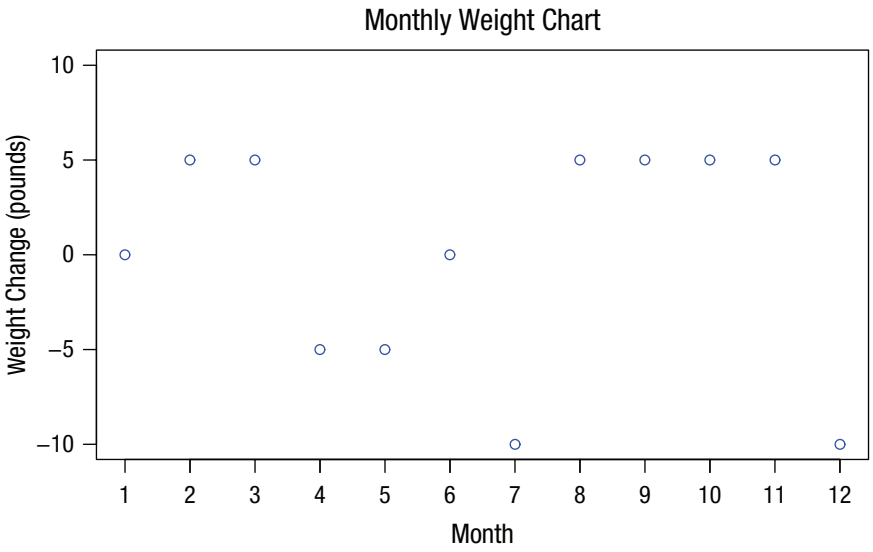


Fig. 4. Sample dot-plot stimulus shown to participants in Study 3b.

Materials and procedure. To generate the stimuli for Study 3b, we first computed every possible set of 12 that included only the values −10, −5, 0, 5, and 10, where the mean of the set equaled zero. We then calculated the imbalance score for the full set of 87 combinations that resulted from this process by subtracting the number of negative values within each set from the number of positive values within each set. The imbalance scores ranged from −4 to 4. We then randomly selected 40 of the combinations, with at least one set for each imbalance score. The 40 combinations were then used as the data to generate 40 dot plots. The plots depicted an individual's weight change per month over a series of 12 months (see Fig. 4).

Each participant viewed a random selection of 10 of the 40 graphs during the study. He or she viewed the menus one at a time and for each was asked, "Based on your immediate response, on average, how much weight change has taken place over the last 12 months?" Answers were made on a scale from −10 to 10.

Results

The imbalance score for each plot corresponded to the number of weight-gain months minus the number of weight-loss months. A linear mixed-effects model, including fixed effects for imbalance, mode, kurtosis, standard deviation, and medium, plus random intercepts for subjects and the by-subject effect of imbalance, found the imbalance score to be the strongest predictor of participants' estimates of the average, β = 0.19, *SE* = 0.07, *b* = 0.44, *SE* = 0.17, bootstrapped 95% CI = [0.07, 0.81], *p* = .009. Standard deviation was also a significant predictor, β = 0.10, *SE* = 0.05, *b* = 0.24, *SE* = 0.12, bootstrapped 95% CI = [0.02, 0.47], *p* = .04. This result suggests that making each data point observable does not counteract the bias. We found that people impose binary categories onto the data they observe, distorting their summary estimates.

Study 4

In the previous studies, the imbalance score and the mode both varied. In Study 4, we found additional evidence for the binary bias by holding the mode constant and varying only the imbalance score.

Method

Participants. Eighty-one participants (43 male; age: *M* = 33.90 years, *SD* = 11.30) from the United States completed the study through Amazon Mechanical Turk.

Materials and procedure. In Study 4, we used the set of histograms with a mean of 3.00 from Study 2 but increased the height of the "3" bar so that it was higher (by at least 20%) than the next tallest bar. The procedure was otherwise identical to that of Study 2a. If people tend to be insensitive to the relative strength of evidence, then even when the mode is held fixed, the imbalance score should continue to predict participants' estimates.

Results

A linear mixed-effects regression with imbalance, kurtosis, and standard deviation as fixed effects, plus random intercepts for subjects and slopes for the by-subject influence of imbalance, predicted participants' mean

estimates. Mode and median were not included in the model because all distributions had a value of 3 for both measures of central tendency. Consistent with the binary bias, results showed that the imbalance score was the only significant predictor of mean estimates, $\beta = 0.33$, $SE = 0.05$, $b = 0.01$, $SE = 0.002$, bootstrapped 95% CI = [.01, .01], $p < .001$. Study 4 differentiates the binary bias from the alternative account of salience (Tsetsos et al., 2012). Even when the mode of distribution corresponds with the correct answer, the binary bias leads to persistent inaccuracy.

Study 5

We next continued to explore the breadth of the binary bias by testing whether it alters judgments in a separate, ecologically valid context: estimating the average price of items on a menu. Unlike in the previous demonstrations, there was no conceptual midpoint, simply a list of prices ranging from \$10 to \$20. We predicted that the binary bias would extend to listed raw data because people would impose an intuitive midpoint on the range of values to simplify data aggregation.

Method

Participants. Eighty participants (54 male; age: M = 33.03 years, SD = 9.89) from the United States completed the study through Amazon Mechanical Turk.

Materials and procedure. To generate prices for the menus in Study 5, we first computed every possible combination of 10 integer prices between \$10 and \$20 with a mean of \$15. This full set of combinations had a range of imbalance scores from −6 to 6. We then randomly selected two combinations for each of the 13 possible imbalance scores to form a stimuli set of 26 menus (see Fig. 5).

Each participant viewed a random selection of 15 of the 26 menus during the study. They viewed the menus one at a time and for each were instructed, "Based on your immediate judgment, please estimate the average price of all items on this menu." After answering from \$10 to \$20, they were then asked, "How would you describe the price range of this restaurant?" (from 0, *very cheap*, to 100, *very expensive*).

Results

The imbalance score for each menu was calculated by subtracting the number of items that cost more than \$15 from the number of items that cost less than \$15. Linear mixed-effects regression models with fixed effects for imbalance, kurtosis, standard deviation, and mode, plus random intercepts for subjects and slopes for the by-subject influence of imbalance, predicted participants' price estimates and expensiveness judgments. Median was not included as a fixed effect because of its strong correlation with mode ($r = .90$, VIF > 4). The imbalance scores of the menus significantly predicted participants' price estimates, $\beta = 0.12$, $SE = 0.04$, $b = 0.03$, $SE = 0.01$, bootstrapped 95% CI = [0.005, 0.06], $p = .03$, but not expensiveness judgments, $\beta = 0.03$, $SE = 0.02$, $b = 0.14$, $SE = 0.12$, bootstrapped 95% CI = [−0.05, 0.37], $p = .25$. Though the direction of the effect of imbalance on expensiveness judgments was in the predicted direction, it may not have reached significance because the expensiveness measure was always the second question for each item and was not attended to as carefully as the price estimates. Additionally, participants may have had difficulty mapping prices onto the expensiveness scale.

The result for mean estimates further suggests that the binary bias is a domain-general strategy for summarizing data. Here, we found that for listed data, participants intuitively divided the items in two and combined them while failing to accurately account for the extremity of each individual price.

Menu Item	Price	Menu Item	Price
Grilled Cheese	\$ 12	Grilled Cheese	\$ 10
Chicken Sandwich	\$ 12	Chicken Sandwich	\$ 10
Hamburger	\$ 14	Hamburger	\$ 16
Cheeseburger	\$ 14	Cheeseburger	\$ 16
Meatball Subs	\$ 14	Meatball Subs	\$ 16
Philly Cheese Steak	\$ 14	Philly Cheese Steak	\$ 16
Mushroom Flatbread	\$ 14	Mushroom Flatbread	\$ 16
Mac and Cheese	\$ 17	Mac and Cheese	\$ 16
Steak and Eggs	\$ 19	Steak and Eggs	\$ 17
Cheese Quesadillas	\$ 20	Cheese Quesadillas	\$ 17

Fig. 5. Sample stimuli items from Study 5. The left-hand menu has an imbalance score of 4, and the right-hand menu has an imbalance score of −6. The true mean of both menus equals \$15.

Study 6

We next examined the relevance of the binary bias to another type of judgment: public-policy decision making.

Method

Participants. Eighty participants (38 male; age: M = 37.50 years, SD = 12.28) from the United States completed the study through Amazon Mechanical Turk.

Materials and procedure. In Study 6, we used pairs of vertical bar charts to depict the carbon dioxide output of two factories. All of the bar charts had an identical mean, but the side of the midpoint with taller bars varied (top heavy vs. bottom heavy). The stimuli for Study 6 were adapted from the set of stimuli with a mean of 3.00 from Study 2. The five distributions with the lowest ratings in Study 2 were randomly matched with the five distributions with the highest ratings. Instead of the restaurant cover story, each distribution was framed as the amount of carbon dioxide being released by a factory. Participants were told, "A government agency wants to cut down on pollution. It must send inspectors to a factory based on the factory's carbon dioxide output over the last 100 months. Which of the two factories should be inspected?" Participants then viewed a pair of distributions. For each distribution, the x-axis was numbered from 1 to 5 and labeled "Millions of Tons of Carbon Dioxide Released," and the y-axis was numbered 1 to 100 in increments of 20 and labeled "Total Months." Each participant viewed all five pairs of distributions.

Results

A Cochran's Q test showed that across the five pairs of graphs, participants chose the high-imbalance graphs more often than chance (M = 3.09, SD = 1.56, 95% CI = [2.75, 3.43]), $Q(4) = 18.66$, $p < .001$. While Study 5 showed that the binary bias influenced estimates of the average, it did not find evidence for an effect on higher-level judgments about expensiveness. The current results show a context in which the binary bias clearly affects high-level processes. This preference for the high-imbalance graph demonstrates another way in which the binary bias goes beyond statistical estimates and affects decision making.

General Discussion

Our studies demonstrate a pervasive bias to treat evidence as binary. Summaries of data are systematically distorted because of a failure to properly weigh the strength of a given piece of evidence and instead evaluate it in an all-or-none manner. The errors presented in these studies are not due to complete misinterpretations of graphs, as sometimes occurs when laypeople misread a graph's entire message (Vekiri, 2002). Instead, they reflect more subtle errors that might influence even the most sophisticated researcher. These errors are not due to any particular visual feature of data visualization but occur even when people are considering information that is not explicitly statistical or visual in any way, suggesting that the error is a domain-general cognitive illusion.

Categorical thinking is not the only factor influencing information integration. In fact, the current studies show that salient data points (operationalized as the mode), kurtosis, and standard deviation independently affect people's summary judgments. Study 1a suggested that the binary bias was a particularly strong factor, as measured against other previously documented order effects such as primacy and recency. However, other cognitive processes need to be considered when developing a full theory of how people summarize conflicting data.

Our studies raise the question of how binary processing takes place. Perhaps the data are encoded categorically and no continuous information is retained. Alternatively, the binary bias may result from making a judgment from a relatively veridical representation. Or continuous information might help identify a midpoint and then no longer be stored. These process-level details suggest directions for future research.

There is a strong computational rationale for the binary bias. By treating all entities in a category (e.g., a particular speech sound) in the same manner, the hearer codes larger scale patterns between categories, a digitization of information that provides powerful processing economies (Harnad, 1987). Treating gradients of information as binary may provide compelling cognitive efficiencies even when it leads to distortions. The root of the binary bias may lie in behavioral control. Many critical behavioral outputs such as fight versus flight and sustenance versus poison are go/no-go (binary) decisions. These survival-relevant processes may shape the sorts of lower-level judgments made in the current studies.

Certain factors may mitigate the bias. Highly numerate individuals have less-distorted probability functions (Patalano, Saltiel, Machlin, & Barth, 2015) and may be less susceptible to the binary bias. Data presentation could also influence how evidence is interpreted. For example, dividing the bins of data differently could change a distribution's imbalance and, thus, its interpretation. Furthermore, contexts with more categories (e.g., average temperature across four seasons) might affect how people bin information.

Cognitive shortcuts allow us to process an otherwise overwhelming amount of information. Here, we

demonstrated that binary bias affects how data are understood and alters decision making. This bias affects not only explicit judgments but also how implicit attitudes are updated on the basis of new evidence. Thus, the binary bias appears to be a pervasive aspect of cognition with extensive real-world implications.

Action Editor

Leaf Van Boven served as action editor for this article.

Author Contributions

M. Fisher developed the study concept, and both authors contributed to the study design. Testing and data collection were performed by M. Fisher. M. Fisher analyzed and interpreted the data under the supervision of F. C. Keil. M. Fisher drafted the manuscript, and F. C. Keil provided critical revisions. Both authors approved the final version of the manuscript for submission.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797618792256

Open Practices

All data and materials have been made publicly available via the Open Science Framework and can be accessed at https://osf.io/kz53g/ and https://osf.io/fvduy/, respectively. Materials have not been made publicly available. The design and analysis plans for the studies were not preregistered. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797618792256. This article has received the badges for Open Data and Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

References

Anderson, N. H. (1981). *Foundations of information integration theory*. San Diego, CA: Academic Press.

Banaji, M. R., & Heiphetz, L. (2010). Attitudes. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (5th ed., Vol. 1, pp. 353–393). Hoboken, NJ: Wiley.

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2014). Fitting linear mixed-effects models using lme4. *arXiv*, Article 1406.5823. Retrieved from https://arxiv.org/pdf/1406.5823.pdf

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). lme4: Linear mixed-effects models using 'Eigen' and S4 (R package Version 1.1-12). Retrieved from https://CRAN.R-project.org/package=lme4

Beach, L. R., & Swenson, R. G. (1966). Intuitive estimation of means. *Psychonomic Science*, *5*, 161–162.

Betsch, T., Kaufmann, M., Lindow, F., Plessner, H., & Hoffmann, K. (2006). Different principles of information integration in implicit and explicit attitude formation. *European Journal of Social Psychology*, *36*, 887–905.

Betsch, T., Plessner, H., Schwieren, C., & Gütig, R. (2001). I like it but I don't know why: A value-account approach to implicit attitude formation. *Personality and Social Psychology Bulletin*, *27*, 242–253.

Corneille, O., & Judd, C. M. (1999). Accentuation and sensitization effects in the categorization of multifaceted stimuli. *Journal of Personality and Social Psychology*, *77*, 927–941.

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2011). Asymmetries in predictive and diagnostic reasoning. *Journal of Experimental Psychology: General*, *140*, 168–185.

Fisher, M., Newman, G. E., & Dhar, R. (2018). Seeing stars: How the binary bias distorts the interpretation of customer ratings. *Journal of Consumer Research*, *45*, 471–489. doi:10.1093/jcr/ucy017

Fleming, S. M., Maloney, L. T., & Daw, N. D. (2013). The irrationality of categorical perception. *The Journal of Neuroscience*, *33*, 19060–19070.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*, 650–669.

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

Goldstein, D. G., & Rothschild, D. (2014). Lay understanding of probability distributions. *Judgment and Decision Making*, *9*, 1–14.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, *24*, 411–435.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*, 767–773.

Harnad, S. (1987). Psychophysical and cognitive aspects of categorical perception: A critical overview. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 1–28). New York, NY: Cambridge University Press.

Henson, R. N. (1998). Short-term memory for serial order: The start-end model. *Cognitive Psychology*, *36*, 73–137.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*, 534–539.

Ibrekk, H., & Morgan, M. G. (1987). Graphical communication of uncertain quantities to nontechnical people. *Risk Analysis*, *7*, 519–529.

Isaac, M. S., & Schindler, R. M. (2013). The top-ten effect: Consumers' subjective categorization of ranked lists. *Journal of Consumer Research*, *40*, 1181–1202.

Johnson, S. G. B., Merchant, T., & Keil, F. C. (2015). Predictions from uncertain beliefs. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th*

Annual Conference of the Cognitive Science Society (pp. 1003–1008). Austin, TX: Cognitive Science Society.

Kahneman, D., Fredrickson, B. L., Schreiber, C. A., & Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science*, *4*, 401–405.

Kvam, P. D., & Pleskac, T. J. (2016). Strength and weight: The determinants of choice and confidence. *Cognition*, *152*, 170–180.

Manis, M., Gleason, T. C., & Dawes, R. M. (1966). The evaluation of complex social stimuli. *Journal of Personality and Social Psychology*, *3*, 404–419.

McShane, B. B., & Gal, D. (2015). Blinding us to the obvious? The effect of statistical training on the evaluation of evidence. *Management Science*, *62*, 1707–1718.

Murphy, G. L., & Ross, B. H. (1994). Predictions from uncertain categorizations. *Cognitive Psychology*, *27*, 148–193.

Patalano, A. L., Saltiel, J. R., Machlin, L., & Barth, H. (2015). The role of numeracy and approximate number system acuity in predicting value and probability distortion. *Psychonomic Bulletin & Review*, *22*, 1820–1829.

Peters, R. G., & Bijmolt, T. H. (1997). Consumer memory for television advertising: A field study of duration, serial position, and competition effects. *Journal of Consumer Research*, *23*, 362–372.

Peterson, C., & Miller, A. (1964). Mode, median, and mean as optimal strategies. *Journal of Experimental Psychology*, *68*, 363–367.

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, *68*, 29–46.

Rosenbaum, M. E., & Levin, I. P. (1968). Impression formation as a function of source credibility and order of presentation of contradictory information. *Journal of Personality and Social Psychology*, *10*, 167–174.

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology*, *8*, 481–520.

Simkin, D., & Hastie, R. (1987). An information-processing analysis of graph perception. *Journal of the American Statistical Association*, *82*, 454–465.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.

Soo, K., & Rottman, B. M. (2018). Causal strength induction from time series data. *Journal of Experimental Psychology: General*, *147*, 485–513.

Spencer, J. (1961). Estimating averages. *Ergonomics*, *4*, 317–328.

Spiegelhalter, D., Pearson, M., & Short, I. (2011). Visualizing uncertainty about the future. *Science*, *333*, 1393–1400.

Tsetsos, K., Chater, N., & Usher, M. (2012). Salience driven value integration explains decision biases and preference reversal. *Proceedings of the National Academy of Sciences, USA*, *109*, 9659–9664.

Vekiri, I. (2002). What is the value of graphical displays in learning? *Educational Psychology Review*, *14*, 261–312.