Original Article

QJEP

Quarterly Journal of Experimental Psychology
2018, Vol. 71(12) 2506–2519

DOI: 10.1177/1747021817744546
qjep.sagepub.com
SAGE

Biasing and debiasing health decisions with bar graphs: Costs and benefits of graph literacy

Yasmina Okan[1,2], Rocio Garcia-Retamero[2,3], Edward T Cokely[3,4] and Antonio Maldonado[2]

Abstract

Bar graphs can improve risk communication in medicine and health. Unfortunately, recent research has revealed that bar graphs are associated with a robust bias that can lead to systematic judgement and decision-making errors. When people view bar graphs representing means, they tend to believe that data points located within bars are more likely to be part of the underlying distributions than equidistant points outside bars. In three experiments, we investigated potential consequences, key cognitive mechanisms, and generalisability of the *within-the-bar bias* in the medical domain. We also investigated the effectiveness of different interventions to reduce the effect of this bias and protect people from errors. Results revealed that the within-the-bar bias systematically affected participants' judgements and decisions concerning treatments for controlling blood glucose, as well as their interpretations of ecological graphs designed to guide health policy decisions. Interestingly, individuals with higher graph literacy showed the largest biases. However, the use of dot plots to replace bars improved the accuracy of interpretations. Perceptual mechanisms underlying the within-the-bar bias and prescriptive implications for graph design are discussed.

Keywords

Graph comprehension; graph design; medical decision-making; graph literacy; risk communication

Received: 29 April 2016; revised: 24 October 2017; accepted: 25 October 2017

Visual displays play an increasingly important role in modern societies, facilitating the communication of complicated information in medicine, economics, weather, climate, and politics (Ancker, Senathirajah, Kukafka, & Starren, 2006; Garcia-Retamero & Cokely, 2013, 2017; Spiegelhalter, Pearson, & Short, 2011). Unfortunately, graphical communication can also cause judgement and decision-making errors. For example, when people are shown a bar graph representing a mean and are asked to judge the likelihood that a data point is part of its underlying distribution, they often believe that the likelihood is larger for points located within the bars than for equidistant points located outside the bars. This tendency, called the "within-the-bar bias" (Newman & Scholl, 2012), is thought to occur because bars are unique visual objects defined by the closure of their boundaries, which originate from one particular axis. Consequently, people's attention is drawn to the region within the bar, such that it takes precedence over regions outside the bar.

Newman and Scholl (2012) demonstrated that the within-the-bar bias affects not only judgements concerning the likelihood of different data points but also decisions made on the basis of bar graphs. They asked participants to imagine they were the CEO of a large car tyre manufacturer and presented them with information concerning the tensile strength of tyres. Participants were told that the mean tensile strength of tested tyres was zero, and that zero was the ideal value for safety. No objective reasons were provided to either increase or decrease the tensile strength of the tyres.

[1]Centre for Decision Research, Leeds University Business School, University of Leeds, Leeds, UK
[2]Department of Experimental Psychology, University of Granada, Granada, Spain
[3]Center for Adaptive Behavior and Cognition (ABC), Max Planck Institute for Human Development, Berlin, Germany
[4]National Institute for Risk & Resilience and Department of Psychology, The University of Oklahoma, Norman, OK, USA

Corresponding author:
Yasmina Okan, Centre for Decision Research, Leeds University Business School, University of Leeds, Charles Thackrah Building, Leeds LS2 9LB, UK.
Email: y.okan@leeds.ac.uk

However, participants who viewed the value of zero represented in a graph where the bar originated from a lower *x* axis (i.e., situated below the mean) often preferred to increase the tensile strength. In contrast, those who viewed this value in a graph where the bar originated from an upper *x* axis (i.e., situated above the mean) often preferred to decrease the tensile strength.

Here, we report three experiments mapping key aspects of the generalisability and mechanisms of the within-the-bar bias. Our central aims in this article were threefold. First, we sought to investigate the extent to which the within-the-bar bias extends to more common health and medical treatment decisions. Second, we aimed to investigate the relations between the bias and a relevant risk literacy skill, namely, graph literacy. Graph literacy refers to the ability to understand and evaluate graphically presented information, and includes general knowledge about making inferences from different graphic formats (Freedman & Shah, 2002; Galesic & Garcia-Retamero, 2011; Kutner, Greenberg, Jin, & Paulsen, 2006). Research suggests that this skill may be a particularly relevant factor in the within-the-bar bias. As compared to less graph literate individuals, more graph literate ones often extract more complex knowledge from line graphs (Maichle, 1994) and more accurately interpret bar graphs depicting interactions (Shah & Freedman, 2011). Graph literacy also robustly predicts the degree to which various users are likely to attend to and integrate decision-relevant information in titles of graphs, axes labels, and scales. In addition, graph literacy predicts lower reliance on salient but not-necessarily diagnostic spatial features during graph interpretation (e.g., heights of bars; Okan, Galesic, & Garcia-Retamero, 2016; Okan, Garcia-Retamero, Galesic, & Cokely, 2012). Accordingly, graph literacy might moderate the within-the-bar bias.

Finally, we investigated the effectiveness of different interventions aimed at reducing the effect of the within-the-bar bias. Specifically, we examined the effects of adding error bars that can emphasise that values from the underlying distributions may come from both below and above the mean (Experiments 1 and 2). We also estimated the relative influence of using dot plots instead of asymmetric bars (Experiment 3). Data corresponding to all experiments and screenshots reflecting all materials viewed by participants can be found in Supplementary Materials.

Experiment 1

We first investigated the effect of the within-the-bar bias on medical decisions by examining participants' preferences for treatments that alter their blood glucose levels. We manipulated whether bars in graphs originated from a lower versus an upper *x* axis, as well as whether graphs contained error bars. To the extent that participants' preferences are affected by the within-the-bar bias, those who receive their blood test results in a bar graph originating from a lower *x* axis (see Figure 1a and c) should seek to increase their blood glucose levels, even if the information gives them no compelling reason to do so. In contrast, those presented with a bar graph descending from an upper *x* axis (see Figure 1b and d) should prefer a treatment that decreases their blood glucose levels. We further predicted that the within-the-bar bias would be moderated by graph literacy, as this skill is generally associated with more skilled decision-making processes, including lower reliance on salient spatial features in graphs (e.g., heights of bars). As a result, higher graph literacy often leads to more accurate graph interpretations and decisions (Okan et al., 2016, 2012; see also Cokely et al., 2018). Finally, we predicted that error bars would reduce the bias particularly among more graph literate viewers, who should be more likely to have the requisite knowledge to effectively interpret and reason about the information conveyed by the error bars.

Method

Participants. Participants were 458 undergraduate students from the University of Granada (307 female), aged 17 to 60 years (lower quartile = 18, median = 19, upper quartile = 22; skewness = 4.38). Two participants did not provide demographic details.

Materials and procedure. The questionnaire was administered in the laboratory of the University of Granada. All materials were implemented as an electronic survey in Unipark (www.unipark.de). As part of another study, the survey first included 30 min of unrelated tasks concerning medical risks, which were followed by the current 15−20 min study (i.e., about 50 min total study time).[1] In this study, all participants were presented with a hypothetical scenario in which they received their blood glucose levels from the previous week. The information was structured building on Newman and Scholl's (2012) vignettes. Participants were informed that a previous measurement of their blood glucose (at the start of the week) had been ideal (120 mg/dL); however, since the start of the week, the last 30 blood tests indicated that their blood glucose levels had varied between −20 and +20 in percentage change. Participants were then reminded that deviation from ideal levels could lead to a high risk of severe health consequences, and that blood glucose levels typically vary throughout the day (e.g., dependent on one's last meal). Participants were further informed that their average percentage change throughout the week was zero.

Participants were randomly assigned into one of five experimental conditions. In the *numerical* (control)

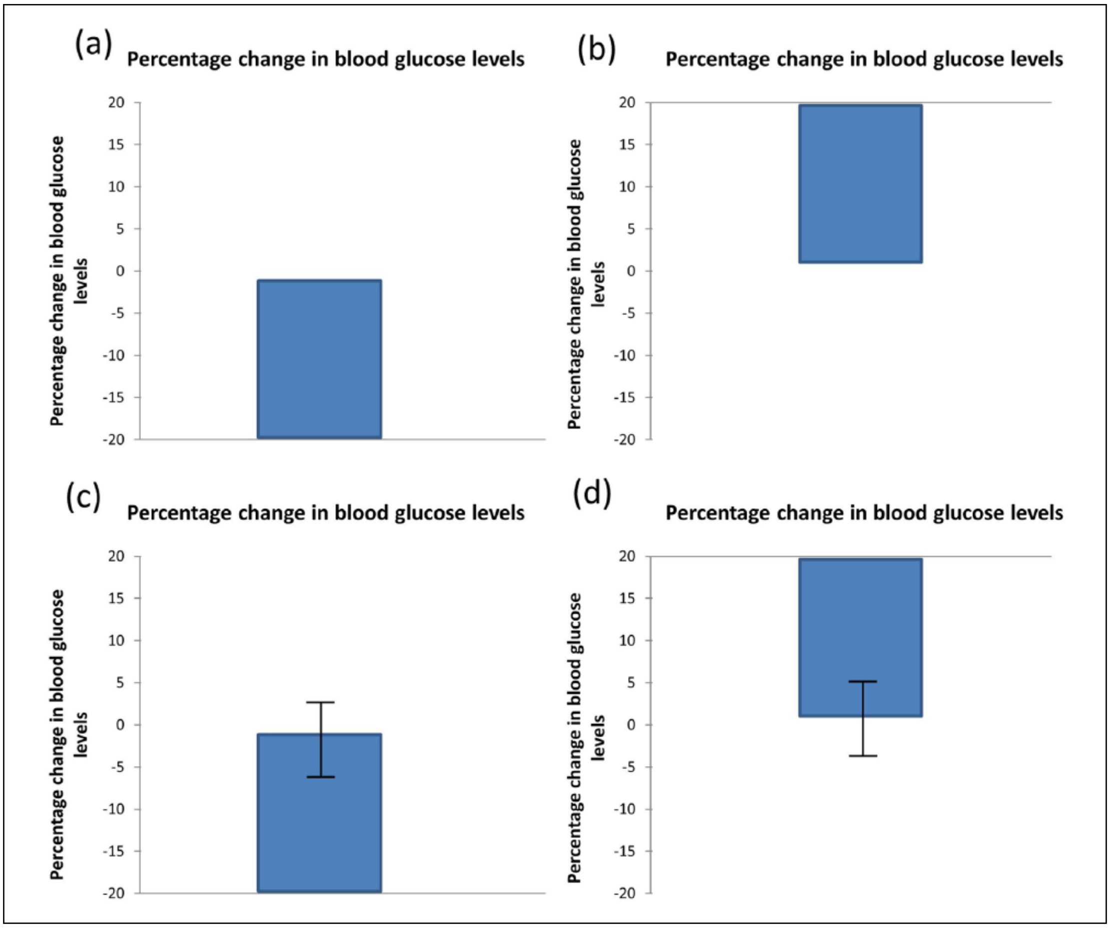

Figure 1. Graphs viewed by participants in Experiments 1 and 2 in the (a) rising, (b) falling, (c) rising with error bars, and (d) falling with error bars conditions.
In Experiment 2, the *y*-axis scale ranged from −40 to +40, and values increased by increments of 10 points.

condition, participants ($n=90$) were presented only with a text containing the numerical information. In the remaining conditions, they were presented with both the numerical information in text and a bar graph depicting this information, which appeared immediately below the text. Participants were informed that the graph showed the average percentage change for the 30 measurements of their blood glucose levels. Bar graphs were constructed following Newman and Scholl (2012). Specifically, in the *rising* condition ($n=91$), the graph displayed a bar rising from a lower *x* axis (see Figure 1a), whereas in the *falling* condition ($n=89$), the bar instead descended from an upper *x* axis (see Figure 1b). Graphs in *rising with error bars* and *falling with error bars* conditions ($n=93$ and $n=95$, respectively) were identical to those in the first two conditions, with the exception that they included bidirectional error bars (see Figure 1c and d). In all cases, the *y*-axis scale ranged from −20 to +20.

Participants were instructed that, based on the information provided, they could choose to follow a treatment that would either slightly increase or slightly decrease their blood glucose levels. They responded using a slider ranging from "slightly decrease my blood glucose levels" to "slightly increase my blood glucose levels," with a mid-point indicating "neither increase nor decrease my blood glucose levels." The numeric slider values ranged from −50 to 50. Following Newman and Scholl (2012), the participants did not see the numerical values. Time to read the scenario and to answer the decision question was unlimited.

Next, graph literacy was measured using the scale developed by Galesic and Garcia-Retamero (2011), which includes a total of 13 items. Graph literacy scores (lower quartile=8.75, median=10, upper quartile=11; skewness=−.60) did not differ across experimental conditions (numerical: $M=9.51$, standard deviation [SD]$=1.78$; rising: $M=9.56$, $SD=1.71$; falling: $M=9.36$, $SD=1.96$; rising with error bars: $M=10.00$, $SD=2.08$; falling with error bars: $M=9.59$, $SD=2.07$), $F(4, 453)=1.41$, $p=.23$. The experiment ended following basic demographic questions and debriefing.[2]

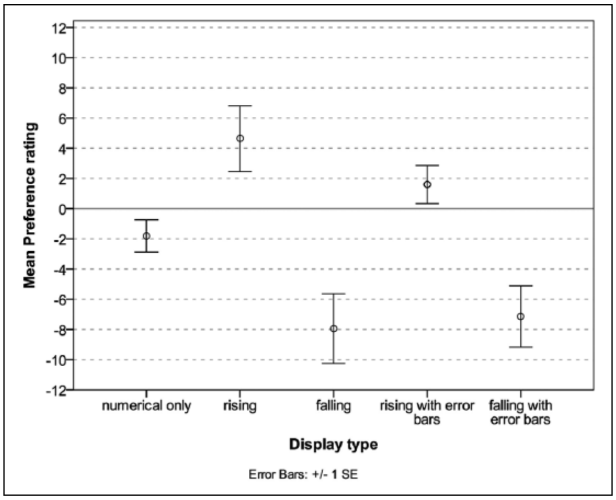

Figure 2. Mean preference ratings by display type in Experiment 1. A mean rating of 0 indicates a preference for maintaining current glucose levels, whereas ratings over and below 0 indicate preference to increase and decrease levels, respectively.

The exact numerical values represented in all figures in the article can be found in Supplementary Materials.

Results

We first examined the effect of the within-the-bar bias on participants' preferences. As predicted and depicted in Figure 2, rising bars led participants to show a preference to increase their blood glucose levels relative to the numerical condition, $t(191.85)=2.95$, $p=.004$, $d=0.38$, 95% confidence interval (CI) = [0.12, 0.63], whereas falling bars resulted in a preference to decrease blood glucose levels relative to the numerical condition, $t(216.85)=4.08$, $p<.001$, $d=0.52$, 95% CI = [0.27, 0.78].

Next, we estimated the extent to which graph literacy moderated the within-the-bar bias, as well as the degree to which error bars reduced the bias. We also examined whether any effect of error bars was stronger among more graph literate individuals. To this end, we computed bias scores by reversing the sign of preference ratings for conditions with falling bars, for comparability with conditions with rising bars. Thus, positive values indicated a preference in the direction expected according to the bias, whereas negative values indicated a preference in the opposite direction. We then constructed a linear regression model predicting bias scores (skewness = .33) from graph literacy scores, the presence of error bars (coded as +1 and −1 for conditions with vs without error bars, respectively), and the interaction between these two factors. Graph literacy scores were mean centred prior to computing the interaction term in this and all other models reported below.[3]

The linear regression model was not a reliable predictor of bias, $R^2=.005$, $F(3, 364)=0.60$, $p=.61$, such that none of the predictors were associated with bias scores (graph literacy: $\beta=-.01$, $t=0.10$, $p=.92$; error bars: $\beta=-.05$, $t=0.93$, $p=.36$; interaction term: $\beta=-.05$, $t=0.93$, $p=.35$). These results suggest that the magnitude of the within-the-bar bias is not a robust function of graph literacy given the current task parameters. Moreover, there was no strong or clear effect of error bars on bias reduction, although a non-significant trend in the expected direction was observed for graphs with rising bars. That is, bias scores were numerically (if not significantly) smaller when error bars were present (see Figure 2), $d=0.18$, 95% CI = [−0.11, 47]. Finally, exploratory analyses also revealed that the bias was overall larger (albeit only slightly) in falling bars conditions ($M=7.53$, $SD=20.70$) than in rising bars conditions ($M=3.11$, $SD=17.01$), $d=0.23$, 95% CI = [0.03, 0.44].

Discussion

The results of Experiment 1 provide the first evidence that the within-the-bar bias can affect medical treatment decisions. Our findings are consistent with the notion that people often mistakenly infer that data points located within bars are more likely to be part of the underlying distribution than equidistant points outside bars. Moreover, our results suggest that this bias may predispose decision-makers to considerable behavioural risks. In our study, the within-the-bar bias was associated with a moderate, robust preference towards modifying one's blood glucose levels in the absence of justifiable reasons to do so.

The current results also suggest that the magnitude of the within-the-bar bias may not reliably vary as a function of one's graph literacy. Even individuals who were relatively skilled in the interpretation and evaluation of graphical information showed similar levels of vulnerability to the bias as less skilled individuals. This finding is somewhat unexpected in the light of the considerable evidence on the decision quality resilience associated with higher levels of graph literacy (e.g., Okan et al., 2016, 2012). However, there are structural elements of the current experimental design that may help to explain the observed boundary condition. For example, participants in our study could extract relevant information from both the text and the graph. Less graph literate individuals may be less comfortable with graphs, and thus they may have spent more time focusing instead on the numerical and text-based information. This may have attenuated the expression of the bias among such individuals. Moreover, the bar graphs had an unusual configuration and displayed fictional data. This may have prompted less graph literate participants to further shift their attention towards the textual information, and away from the stimuli that is responsible for the bias (i.e., the graphical materials). A stronger bias among individuals with lower graph literacy may only emerge when all participants allocate a similar amount of attention to the graphs.

Finally, our findings also suggest that error bars will not necessarily reduce the within-the-bar bias, although the tendency at the descriptive level was in the expected direction for graphs with rising bars. To further explore potential mechanisms and boundaries of the within-the-bar bias, we conducted a second experiment investigating the effects of error bars. We also examined whether graph literacy affects the bias after equating the degree to which all participants are required to attend to the graph.

Experiment 2

Experiment 2 was designed to address three new questions. First, we sought to determine whether graph literacy would affect the magnitude of the within-the-bar bias when people are required to attend to both the text and the graph. Second, we examined whether the bias extends to a scenario involving a different reference point for initial blood glucose levels. The reference point described in Experiment 1 (120 mg/dL) may have been perceived as high by participants, considering that a fasting glucose level of 126 mg/dL or more is associated with a diagnosis of diabetes (American Diabetes Association, 2012). Moreover, high blood glucose levels might be perceived as having more severe consequences than low blood glucose levels, even though hospital admission rates for the latter cause can be higher in certain populations (Lipska et al., 2014). Thus, in Experiment 2, we used a scenario that described a lower initial reference point (100 mg/dL).

Finally, in Experiment 2, we also estimated the extent to which the within-the-bar bias would affect people's judgements concerning the likelihood that different data points were part of the underlying distribution. Participants' treatment preferences in Experiment 1 were consistent with the assumption that people often believe that a given data point is more likely to be part of the distribution when it is located within the bar than outside the bar. However, we did not assess likelihood judgements directly. Thus, in Experiment 2, we also asked participants to judge the likelihood of two different blood glucose measurements (one below the mean and another one above the mean). We expected that the within-the-bar bias would lead participants presented with a rising bar to judge the measurement below the mean as more likely than the measurement above the mean, as the rising bar encompasses values below this point. Instead, we expected to find the reverse pattern among those presented with a falling bar. That is, the measurement above the mean should be judged as more likely in this case, as the falling bar comprises values above the mean.

Method

Participants. Participants were recruited via Amazon's Mechanical Turk, which provides access to a paid Internet participant panel that has been widely used for behavioural decision-making research (Chandler & Shapiro, 2016; Paolacci & Chandler, 2014). The task was available only to individuals who had an acceptance rate greater than or equal to 95% in previous human intelligence tasks (HITs) on Mechanical Turk, following recommendations to ensure high quality data (Peer, Vosgerau, & Acquisti, 2014). A total of 954 U.S. residents clicked on the link to our study and 822 completed it. Three participants completed the survey after a break and one participant experienced technical problems with the survey. These participants were excluded from our analyses based on a priori criteria to exclude participants who did not complete the survey in one sitting. The final sample included 818 participants (525 women, age range 18-77 years, lower quartile = 26, median = 33, upper quartile = 47; skewness = .78). Nine percent had no more than a high school diploma, 39% had completed up to some college or associate degree, 37% had a bachelor's degree, and 15% had a master's degree or higher. One participant did not indicate his or her educational level. The average completion time was 18 min.[4]

Materials and procedure. The web survey was programmed using Unipark (www.unipark.de). Participants were redirected to the survey after clicking on a link provided in the HIT forum on Mechanical Turk. Materials presented to participants were identical to those in Experiment 1, with the exception that the scenario stated that the value for the measurement taken at the start of the week had been 100 mg/dL and blood glucose levels had varied between −40 and +40 in percentage change. The y-axis scale in graphs ranged from −40 to +40, with values increasing in increments of 10 points (see Supplementary Materials, Figures S8-S11).

Participants were randomly assigned to one of the five experimental conditions used in Experiment 1 (numerical: $n = 172$; rising: $n = 166$; falling: $n = 161$; rising with error bars: $n = 154$; falling with error bars: $n = 165$). However, information was displayed differently, with the aim of ensuring that participants attended the graphs as well as the accompanying text. Specifically, in the numerical only condition, the textual information was first presented alone on one screen. This information was then presented again on the next screen, accompanied by the slider to assess participants' preferences. In all remaining conditions, the textual information was first presented alone on one screen, followed by the graph alone on the next screen. Participants were informed that the graph showed the average percentage change for the 30 measurements of their blood glucose levels, and were instructed to take some time to look at the information represented. Finally, both the textual information and the graph appeared together on the same screen, accompanied by the slider to assess preferences. Participants in all conditions were required to view the text alone for at least 10 s, before they could move onto the next screen. To this end, the Continue button was not visible until 10 s after the screen containing the text had been displayed. In the conditions including graphs, this also applied to the screen displaying the graph alone.[5]

As noted above, in Experiment 2, we also assessed participants' judgements of the likelihood that values above versus below the mean were part of the underlying distribution. The question assessing the perceived likelihood of the value above the mean was as follows: "What do you think is the likelihood that one of your blood glucose level measurements was 120 mg/dL (i.e., an increase of 20% from the measurement taken at the start of the week)?" The question assessing the perceived likelihood of the value below the mean was identical, with the exception that it referred to a measurement of 80 mg/dL (i.e., a decrease of 20% from the measurement taken at the start of the week). Participants responded using a 7-point scale ranging from 1 (*extremely unlikely*) to 7 (*extremely likely*). The order of likelihood ratings was counterbalanced. All remaining aspects of the procedure were identical to that of Experiment 1.[6]

Graph literacy scores (lower quartile = 9, median = 11, upper quartile = 12; skewness = −1.24) did not differ across experimental conditions (numerical: $M = 10.54$, $SD = 1.80$; rising: $M = 10.40$, $SD = 1.96$; falling: $M = 10.50$, $SD = 1.96$; rising with error bars: $M = 10.46$, $SD = 1.94$; falling with error bars: $M = 10.58$, $SD = 1.95$), $F(4, 813) = 0.23$, $p = .92$.

Results

The within-the-bar bias again affected preferences in the expected direction. As can be seen in Figure 3a, rising bars were associated with a preference to increase blood glucose levels relative to the numerical condition, $t(410.56) = 2.45$, $p = .02$, $d = 0.23$, 95% CI = [0.05, 0.42], whereas falling bars instead led participants to prefer to decrease their levels relative to the numerical condition, $t(407.42) = 9.61$, $p < .001$, $d = 0.91$, 95% CI = [0.71, 1.10].

A linear regression, including graph literacy scores, presence of error bars, and the interaction between these factors as predictors of bias scores (computed using the same procedure as in Experiment 1; skewness = .35), explained a small but significant amount of variance, $R^2 = .02$, $F(3, 642) = 3.38$, $p = .02$. In contrast to Experiment 1, in this study, graph literacy scores significantly predicted bias in preference ratings. Interestingly, however, higher scores were related to modest, yet, significantly stronger bias, $\beta = .12$, $t = 3.00$, $p = .003$. Error bars and the interaction term were not significant predictors ($\beta = -.03$, $t = 0.78$, $p = .44$ and $\beta = -.03$, $t = 0.76$, $p = .45$, respectively), although for graphs with rising bars, there was again a non-significant trend in the expected direction (see Figure 3a), $d = 0.13$, 95% CI = [−0.09, 0.35]. In line with Experiment 1, exploratory analyses also revealed that the

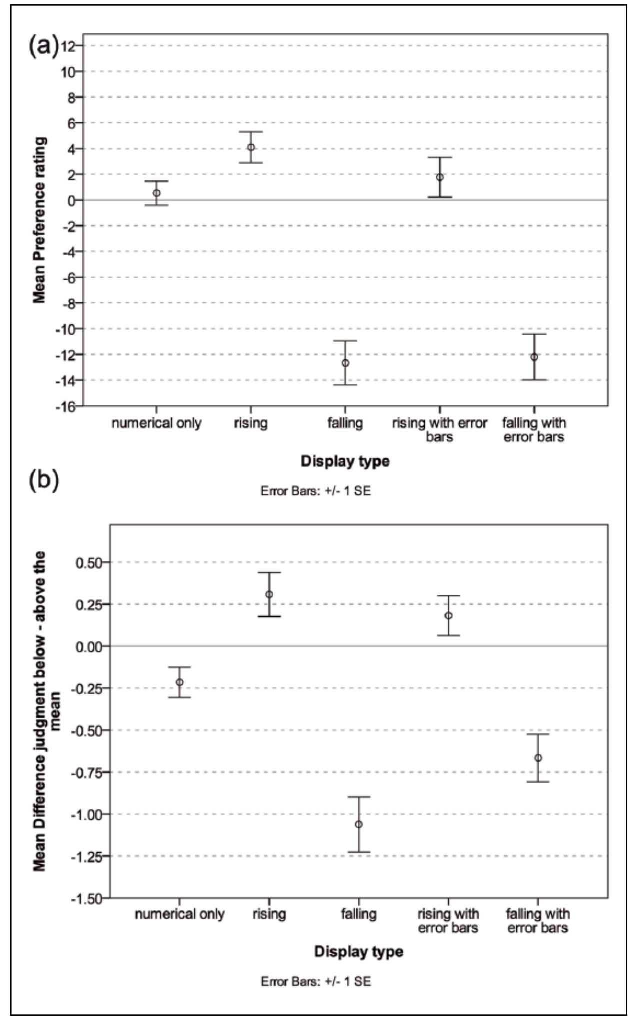

Figure 3. (a) Mean preference ratings by display type in Experiment 2. A mean rating of 0 indicates a preference for maintaining current glucose levels, whereas ratings over and below 0 indicate preference to increase and decrease levels, respectively. (b) Differences between likelihood ratings corresponding to judgements below the mean (80 mg/dL) and above the mean (120 mg/dL) by display type in Experiment 2. The exact numerical values represented in all figures in the article can be found in Supplementary Materials.

bias was overall larger in conditions with falling bars ($M = 12.43$, $SD = 22.29$) versus rising bars ($M = 2.98$, $SD = 17.40$), $d = 0.47$, 95% CI = [0.32, 63].

Next, we examined participants' likelihood judgements. Consistent with the anticipated influence of the within-the-bar bias, in the falling condition, the blood glucose measurement above the mean was judged to be significantly more likely ($M = 5.12$, $SD = 1.59$) than the measurement below the mean ($M = 4.06$, $SD = 1.93$), paired $t(160) = 6.48$, $p < .001$, $d = 0.73$, 95% CI = [0.48, 1.03]. This was also the case in the falling with error bars condition (judgement above: $M = 4.62$, $SD = 1.68$; judgement below: $M = 3.95$, $SD = 1.81$), paired $t(164) = 4.67$, $p < .001$, $d = 0.46$, 95%

CI = [0.21, 0.74]. As expected, this trend reversed in the rising condition, where the measurement above the mean was judged to be less likely (M = 4.20, SD = 1.86) than the measurement below the mean (M = 4.51, SD = 1.81), paired $t(165)$ = 2.34, p = .02, d = 0.26, 95% CI = [0.02, 0.54]. A non-significant trend in the anticipated direction was also observed in the rising with error bars condition (judgement above: M = 4.38, SD = 1.65; judgement below: M = 4.56, SD = 1.56), paired $t(153)$ = 1.53, p = .13, d = 0.17, 95% CI = [−0.08, 0.43].

To quantify the bias in likelihood ratings, for each participant we deducted the rating corresponding to the value above the mean from the rating corresponding to the value below the mean. We then reversed the sign in the falling conditions, for comparability with the rising conditions, and constructed a linear regression model predicting bias in likelihood ratings (skewness = .77) from graph literacy scores, error bars, and the interaction between these factors. This model also explained a small but significant amount of variance, R^2 = .02, $F(3, 642)$ = 4.32, p = .005. Graph literacy scores predicted bias in likelihood ratings, with higher scores again relating to stronger bias, β = .12, t = 3.03, p = .003. As can be seen in Figure 3b, there was again a trend for error bars to reduce bias, although this factor did not reach conventional levels of significance, β = −.07, t = 1.80, p = .07. The interaction term between error bars and graph literacy was also not significant, β = −.03, t = 0.84, p = .40, and presented no evidence of any notable trend. Exploratory analyses again revealed a stronger bias in conditions including falling bars (M = 0.86, SD = 1.97) versus rising bars (M = 0.25, SD = 1.60), d = 0.34, 95% CI = [0.19, 50].

Discussion

Results of Experiment 2 showed that the within-the-bar bias not only affected participants' preferences for different medical treatments but also their judgements concerning their likelihood of having a given blood glucose value. Interestingly, we also found that this bias was *more* marked among more graph literate participants (cf. Okan et al., 2016, 2012). One possible explanation is that, even though all participants were required to allocate a similar amount of attention to the bar graphs overall, less graph literate participants may have attended to a lesser extent to the values on the y axis (see also Okan et al., 2016). Indeed, the within-the-bar bias cannot arise if graph viewers do not encode the values on the y axis, as associations must be established between the region within the bars and the corresponding values on the graph (e.g., values below the mean, for rising bars). Eye-tracking evidence supports this interpretation, as studies have revealed that lower graph literacy is associated with shorter viewing times of conventional features in graphs, such as axes labels or scales (Okan et al., 2016). It is also possible that less graph literate participants were not able to generate a detailed mental model of the bar graph, without which the within-the-bar bias may not arise. Such differences in processing and comprehension capability may have resulted in a reduced susceptibility to this bias among individuals with lower graph literacy.

In Experiment 2, we again found evidence suggesting that error bars may not reliably reduce the bias, although there was a marginally significant difference in the expected direction for likelihood ratings. Given that all participants were required to attend to the graph in this experiment, it seems unlikely that the limited effectiveness of error bars merely reflects that this design feature was neglected. Thus, in Experiment 3, we further evaluated potential boundary conditions by examining a different intervention that theory suggests may be more effective in reducing the within-the-bar bias, namely, the use of dot plots to represent means.

In addition, an important question that remains unanswered is whether the within-the-bar bias is robust enough to affect people's interpretations of graphs that communicate relevant medical or health information to the public. Stimulus materials in Experiments 1 and 2 were designed to foster high internal validity and allow clear theory evaluation. Nevertheless, it remains unclear whether the findings documented may generalise to graphs used in ecological, naturalistic contexts. This question is theoretically and practically relevant because simple graphical displays including bar graphs are increasingly used and recommended to communicate health information to diverse, and often vulnerable, populations facing high-stakes medical decisions (see, for example, Garcia-Retamero & Cokely, 2013; Lipkus, 2007; Trevena et al., 2013).

Experiment 3

Our main goal in Experiment 3 was to estimate the extent to which the within-the-bar bias may affect people's judgements in relation to ecological materials that are more representative of common naturalistic decision-making. Specifically, we turned to the website of the Centers for Disease Control and Prevention (CDC), which features a wide-ranging pool of publicly available graphs summarising results of national healthcare surveys conducted by the U.S. National Center for Health Statistics (NCHS). Such statistical information is explicitly intended to inform and guide actions and policies in the service of benefiting the health and welfare of people in the United States. To the extent that the within-the-bar bias affects interpretations of graphs in this website, such bias could ultimately have an adverse effect on health policy and outcomes. We focused on information concerning the consumption of added sugars among U.S. adults given the implications for preventing obesity and diabetes, and the dramatic increase in the prevalence of these diseases in the last decades (World Health & Organization, 2017a, 2017b).

An additional goal of Experiment 3 was to test the effectiveness of dot plots to reduce any effect of the within-the-bar bias. Dot plots were recommended as an alternative to bar charts by Cleveland (1984) and Cleveland and McGill (1984) based on the notion that they allow for more effective visual decoding of data. Newman and Scholl (2012) also noted that the use of points to represent means instead of asymmetric bars could improve the accuracy of graph interpretations. Dots do not need to be connected to the *x* axis, and they may attract people's attention to a larger extent than the space between the dots and the axis (Godau, Vogelgesang, & Gaschler, 2016). Thus, this kind of display should be less likely to trigger systematic biases in people's judgements of the likelihood of different data points. However, to our knowledge, this prediction has not yet been tested. In Experiment 3, we examined this issue by comparing people's interpretations of a bar graph selected from the CDC website versus an alternative version of the graph in which bars were replaced by simple dots (see Figure 4). In line with previous experiments, we also examined people's interpretations of data when presented with numerical information only (as a control condition), which in this case was displayed in a tabular format.

As in Experiments 1 and 2, we expected that participants presented with the bar graph would be affected by the within-the-bar bias. As the selected graph contained rising bars, we expected that participants would judge values below the depicted means as more likely than equidistant values above the means. We also predicted that dot plots would contribute to reduce or eliminate the bias.

Finally, in Experiment 3, we also examined participants' evaluations of the materials. Understanding how different types of displays are evaluated is important because people may not be motivated to attend to, or take actions regarding, graphs that they dislike (Ancker et al., 2006; Okan, Stone, & Bruine de Bruin, 2017; Stone, Bruine de Bruin, Wilkins, Boker, & MacDonald Gibson, 2017). There is evidence that simple bar graphs are on some occasions preferred over other types of graphs, such as line graphs, icon arrays, and survival curves (Fortin, Hirota, Bond, O'Connor, & Col, 2001). There is also evidence that bar graphs can signal more scientific credibility than verbal descriptions, enhancing people's beliefs in the efficacy of products (Tal & Wansink, 2016). It is possible that bar graphs will be associated with more positive user evaluations than less widespread formats, such as dot plots, despite the potential of the former type of graph to bias people's interpretations and decisions.

Participants

Participants were recruited following the same procedure as in Experiment 2. A total of 672 U.S. residents clicked on the link to our study and 612 completed it. One participant indicated that his or her age was 5 years and was thus excluded from subsequent analyses. The final sample included 611 participants (352 women, age range 18-77 years, lower quartile = 27, median = 33, upper quartile = 44; skewness = .89). Eight percent had no more than a high school diploma, 37% had completed up to some college or associate degree, 41% had a bachelor's degree, and 14% had a Master's degree or higher. One participant did not indicate his or her educational level. The average completion time was 15 min.

Materials and procedure

The procedure used to host the web survey was identical to that used in Experiment 2. Participants were informed that they would view data from the National Health and Nutrition Survey concerning the consumption of added sugars among U.S. adults between 2005 and 2010. Participants were further informed that increased consumption of added sugars has been linked to a decrease in intake of essential micronutrients and an increase in body weight. All information was based on that included in the data brief concerning this topic available on the CDC website (Ervin & Ogden, 2013).

Participants were randomly assigned into one of the three experimental conditions. In the *table* (control) condition ($n = 207$), participants were presented with a simple table summarising the data (see Figure 4a). In the *bars* condition, participants ($n = 202$) were presented with the original bar graph taken from the CDC data brief, depicting mean kilocalories from added sugars consumed per day among adults aged 20 years and over, by age group and sex (see Figure 4b). Finally, participants in the *dot plot* condition ($n = 202$) were presented with a redesigned version of the original bar graph, which was identical to the original in all respects, with the exception that bars were replaced by dots (see Figure 4c).

Participants were required to judge the likelihood that an individual in one of the groups represented (a female aged between 20 and 39 years) had consumed a given amount of kilocalories of added sugars, which was either above or below the mean for that group. The question concerning the value above the mean was as follows: "What do you think is the likelihood that a female aged between 20 and 39 consumed around 425 kcal of added sugars on a given day?" The question concerning the value below the mean was identical, with the exception that it enquired about a value of 125 kcal. As can be seen in Figure 4, the average kilocalories consumed by this group was 275, implying that the values enquired about were equidistant to the mean. Participants responded using the same 7-point scale as in Experiment 2, and the order of likelihood ratings was again counterbalanced.[7]

(a) **Table 1. Mean kilocalories from added sugars per day among adults aged 20 and over, by age group and sex: United States, 2005–2010**

	Total	20-39 years	40-59 years	60 years and over
Male	335	397	338	224
Female	239	275	236	182

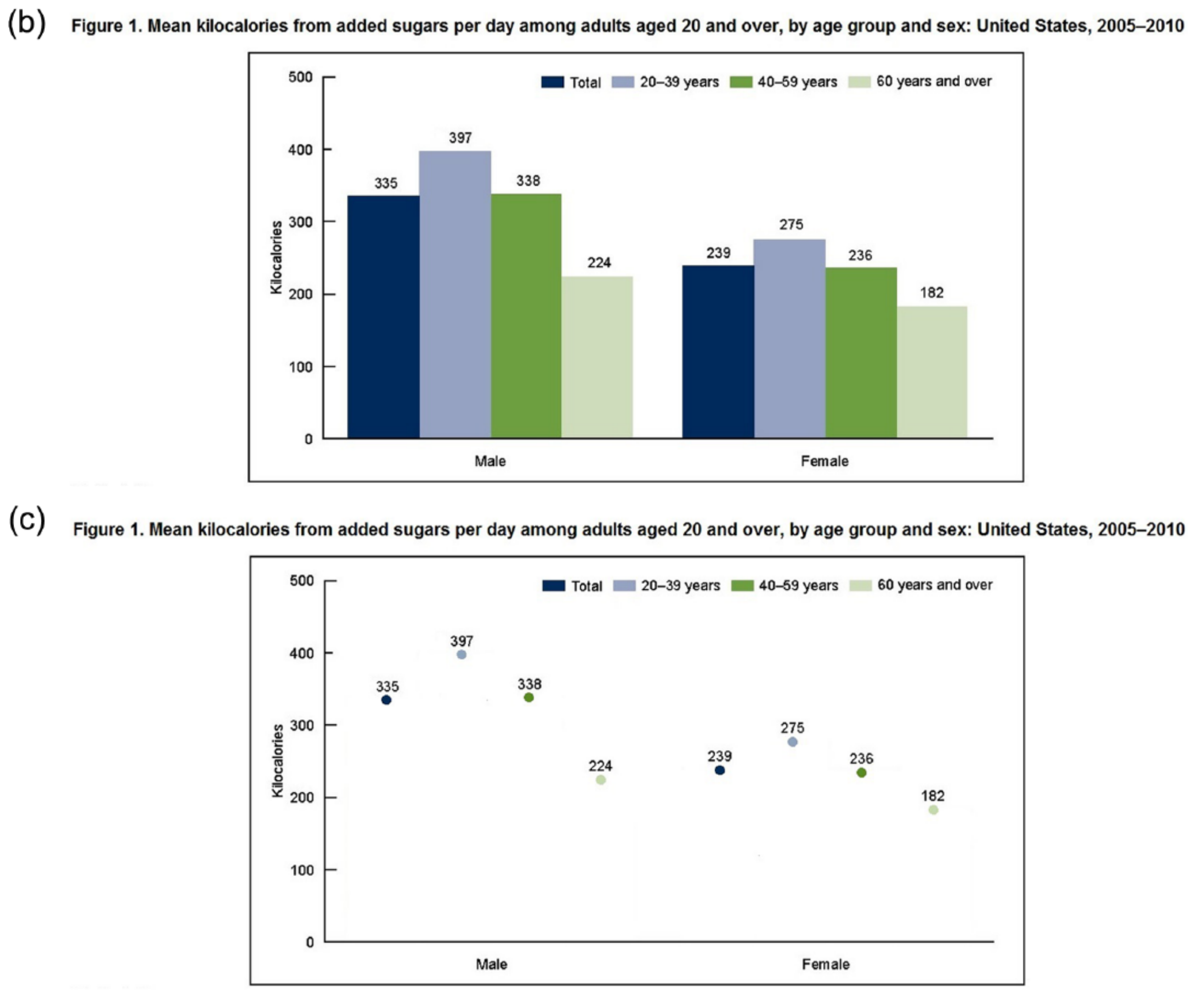

Figure 4. Displays viewed by participants in Experiment 3 in the (a) table, (b) bars, and (c) dot plot conditions (colour figures available online).

The graph presented in the bars condition was taken from the Centers for Disease Prevention and Control (CDC) website (http://www.cdc.gov/nchs/data/databriefs/db122.htm). The original graph contained superscript numbers next to some of the values at the top of the bars to indicate statistically significant differences between the groups. Superscripts were removed to avoid confusion.

User evaluations of the materials were next assessed with three items asking participants to rate how much they liked the way in which the data were presented, how helpful was the table/graph for making decisions regarding the consumption of added sugars, and how much they would trust information represented in a table/graph like the one they viewed, using a scale from 1 to 7 (see Bruine de Bruin, Stone, Gibson, Fischbeck, & Shoraka, 2013 for a similar procedure). We computed a composite measure of user evaluations by averaging participants' responses across all three items (Cronbach's alpha = .86). All remaining aspects of the procedure were identical to that of Experiment 2.

Graph literacy scores (lower quartile = 9, median = 11, upper quartile = 12; skewness = −1.57) again did not differ across experimental conditions (table: M = 10.39, SD = 2.27; bars: M = 10.31, SD = 2.31; dot plot: M = 10.29, SD = 2.20), $F(2, 608) = 0.12$, $p = .89$.

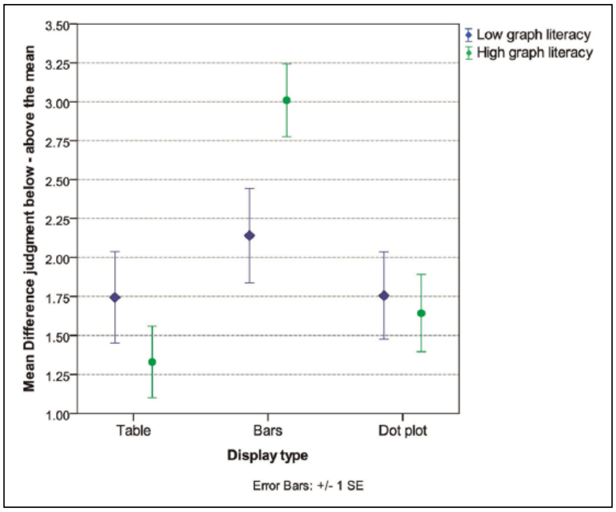

Figure 5. Differences between likelihood ratings corresponding to judgements below the mean (125 kcal) and above the mean (425 kcal) by display type and graph literacy in Experiment 3. In this figure, participants are categorised as low graph literates if they obtained 10 or fewer correct responses (n = 261, mean score = 8.4, SD = 2.1), and as high graph literates if they obtained 11 or more (n = 350, mean score = 11.8, SD = 0.7), according to a median split. However, continuous graph literacy scores are entered in all analyses.
The exact numerical values represented in all figures in the article can be found in Supplementary Materials.

Results

Consistent with previous findings, participants presented with bars judged the value below the mean (M = 5.26, SD = 1.82) to be more likely than the value above the mean (M = 2.62, SD = 1.62), paired $t(201)$ = 14.06, p < .001, d = 1.40, 95% CI = [1.15, 1.62], revealing a large, significant influence of the within-the-bar bias. A tendency in the same direction was also observed among those presented with the table (judgement below: M = 4.25, SD = 2.05; judgement above: M = 3.01, SD = 1.60) and the dot plot (judgement below: M = 4.41, SD = 2.21; judgement above: M = 2.72, SD = 1.67).

To examine the relative difference between ratings concerning values below versus above the mean for the different display types, we constructed a linear regression predicting bias in likelihood ratings (i.e., differences between both judgement types; skewness = −.10) from display type, using dummy coding with the bars condition as the reference category. Graph literacy and the interaction between graph literacy and display type were also included as predictors. This model explained a moderate and significant amount of variance, R^2 = .05, $F(5, 605)$ = 6.57, p < .001. As expected, bias was significantly smaller in the table than in the bars condition, β = −.20, t = 4.42, p < .001, and dot plots significantly reduced the bias, β = −.17, t = 3.66, p < .001 (see Figure 5). Graph literacy scores predicted bias scores with higher graph literacy related to stronger bias, β = .22, t = 3.23, p = .001. The interaction terms between graph literacy and display type were also significant to marginally significant (β = −.13, t = 2.40, p = .02 for bars vs table and β = −.10, t = 1.82, p = .07 for bars vs dot plot). As illustrated in Figure 5, differences between the bars condition versus the table and dot plot conditions were larger among more graph literate individuals. In addition, the correlation between graph literacy and bias was only significant in the bars condition (bars: r = .22, p = .001; table: r = −.01, p = .86; dot plot: r = −.04, p = .58), once again revealing that the within-the-bar bias tended to be larger among more graph literate individuals.

Finally, we examined participants' evaluations of the materials. As anticipated, bar graphs were evaluated more positively (M = 4.95, SD = 1.41) than tables (M = 4.56, SD = 1.46), $t(407)$ = 2.74, p = .01, d = 0.27, 95% CI = [0.08, 0.47] and dot plots (M = 4.58, SD = 1.43), $t(402)$ = 2.65, p = .01, d = 0.26, 95% CI = [0.07, 0.46], despite the notable reduction of bias associated with the latter two display types.

Discussion

In Experiment 3, we replicated and extended findings of Experiments 1 and 2. The within-the-bar bias affected participants' interpretations of ecological graphs concerning current health topics, designed to guide actions relevant to the promotion and maintenance of public health policies. In line with Experiment 2, we also found that the bias was stronger among more graph literate participants.

In Experiment 3, we also documented the first evidence on the effectiveness of dot plots to reduce the within-the-bar bias in a theoretically and practically relevant context. This type of graph markedly reduced the expression of bias, providing additional empirical validation of long-standing recommendations on the benefits of dot plots for improving graph interpretations (Cleveland, 1984; Cleveland & McGill, 1984). Interestingly, and somewhat ironically, bar graphs were evaluated more positively than dot plots and tables. This finding may reflect participants' general familiarity with bar charts, and adds to the increasing body of work showing that people's preferences for different display types may run counter to what is best for their overall performance (Feldman-Stewart, Kocovski, McConnell, Brundage, & Mackillop, 2000; McCaffery et al., 2012; Okan, Garcia-Retamero, Cokely, & Maldonado, 2015; Waters, Weinstein, Colditz, & Emmons, 2006). That said, it is notable that in this study, all types of displays received relatively positive user evaluations. Thus, although not necessarily the most favoured option, dot plots can be a welcome and promising graphical format that promotes more accurate interpretations among users who vary widely in ability and backgrounds.

General discussion

In three experiments, we showed that bar graphs depicting means can systematically result in misinterpretation, thereby biasing people's judgements and causing decision vulnerabilities. Our findings revealed that the within-the-bar bias can affect people's preferences for different medical treatments, as well as inferences about ecological and naturalistic graphs designed to support informed decision-making by governmental agencies. Moreover, in two experiments, we found, ironically, that more graph literate participants may be at greater risk for within-the-bar bias. These results appear particularly noteworthy considering that graph literacy generally is associated with lower risk of various biases and misunderstandings (e.g., Okan et al., 2016, 2012), and given that the use of bar graphs to communicate health-related information is widespread (Garcia-Retamero & Cokely, 2013; McCaffery et al., 2012; Mt-Isa et al., 2013). Nevertheless, the current findings also point to a potentially promising method to overcome the within-the-bar bias, namely, replacing bar graphs with simple dot plots.

Concerning the perceptual mechanisms that give rise to the within-the-bar bias, Newman and Scholl (2012) argued that the bias occurs because bars are unique visual objects defined by the closure of their boundaries, which originate from one particular axis. Relatedly, Peebles (2008) demonstrated that people presented with bar graphs underestimated the distance of target values to the average (represented by a horizontal line parallel to the *x* axis). More recently, Godau et al. (2016) documented converging evidence that people systematically underestimate mean values in graphs with rising bars, independently of the height of bars. Theoretically, visual attention is drawn to the length of bars, which are identified as objects attached to the *x* axis. These accounts converge with our findings to indicate that the within-the-bar bias is likely triggered by basic principles of object perception. Bottom-up factors such as the format of graphs can influence the visual chunks that are created, often driven by Gestalt principles, including proximity, similarity, and connectedness (Ali & Peebles, 2013; Pinker, 1990). Although the visual chunks formed by bars can facilitate tasks such as making discrete comparisons between individual data points (Pinker, 1990) and interpreting interaction data (Ali & Peebles, 2013), they can also lead to systematic misinterpretations of bar graphs.

Cognitive process tracing methodologies, such as eye-tracking and verbal protocol analysis, could be used to shed further light on the role of perceptual and attentional processes underlying the within-the-bar bias. Such methods could also help to map the mechanisms underlying the debiasing effects of dot plots. For instance, eye-tracking methodology could be used to determine whether the dots attract people's attention to a larger extent than the space between dots and the *x* axis (Godau et al., 2016), and the extent to which any attentional differences affect interpretations. Process tracing methods could also help to understand how people perceive and interpret error bars, as well as their relative effectiveness in different contexts (or lack thereof), for different viewers. Future research could also investigate the effect of the within-the-bar bias on representative decisions with real stakes for decision makers, families, organisations, and societies. Finally, future research should assess the robustness of the observed effects across heterogeneous samples in terms of graph literacy and other cognitive, social, and demographic variables. We speculate that the relationship between graph literacy and the bias may often be curvilinear, such that highest graph literacy levels may be associated with a lower bias. That is, we suspect that expert scientists and statisticians will not exhibit a within-the-bar bias and will be more likely to correctly interpret error bars.

Conclusion

The present work provides new evidence that bar graphs depicting means can be associated with systematic biases likely caused by common, basic principles of object perception. We also found that such biases can predispose decision makers to misinterpretations and judgement errors that may have counterproductive and potentially dangerous downstream effects on health-related decision-making. Surprisingly, we also found some of the first evidence that essential risk literacy skills (i.e., graph literacy) may promote rather than reduce decision vulnerability. We suspect these effects may be best characterised as reflecting issues that result from modest but still relatively insufficient skills. That is, high expert level decision-makers may not be affected by the within-the-bar bias, whereas normally sufficient levels of skill may predispose individuals to this and other potentially costly biases.

Due to the perceptual nature of the within-the-bar bias, even bar graphs designed according to principles of effective graph design have the potential to mislead viewers. Although the implications of this failure should not be discounted, we also found that other formats may address this issue. That is, graph designers may be able to use alternative graphical formats (e.g., points or depictions of the distributions) to represent means to good effect, helping reduce decision and interpretational vulnerabilities. Taken together, the present research adds to the increasing body of literature on skilled decision-making and the design of interventions that promote informed decision-making. Our work also contributes to theories on graphical risk communication that aim to predict when and why biases will occur, and how to best design graphs and communications that empower diverse decision-makers facing high-stakes personal, social, and economical decisions.

Acknowledgements

Part of this research was conducted as part of Y.O.'s doctoral dissertation at the University of Granada. The authors thank all members of the committee for their feedback. They are also grateful to Mirta Galesic, Andrés Catena, Wändi Bruine de Bruin, Eric Stone, Catherine Fritz, and two anonymous reviewers for their valuable comments and suggestions.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by a George Bennett Dissertation Fellowship (ref. 0230-1) from the Informed Medical Decisions Foundation (United States) and a grant awarded by the Worldwide Universities Network (Fund for International Research Collaborations) to Y.O. Y.O.'s time was also in part supported by a Population Research Fellowship awarded by Cancer Research UK. Other partial financial support was provided by grants PSI2011-22954 and PSI2014-51842-R awarded by the Ministerio de Economía y Competitividad (Spain), and grants from the U.S. National Science Foundation (SES-1253263), the Gulf Research Program of the National Academies of Sciences, Engineering, and Medicine, the RiskLiteracy.org WebMD-Medscape CME programme SF232838, with additional financial support from the National Institute for Risk & Resilience and the University of Oklahoma. The authors declare independence from these funding agencies in each of the following: design of the study: collection, management, analysis, and interpretation of the data, and preparation of the manuscript.

Supplemental material

The online Supplemental Material is available at: journals.sagepub.com/doi/suppl/10.1177/1747021817744546.

Notes

1. In the unrelated tasks, participants were presented with visual aids (icon arrays) depicting the effectiveness of hypothetical drugs for heart attack prevention. We assessed participants' risk understanding, confidence in their risk estimates, and evaluations of the visual aids. Further details concerning this part of the survey can be found in Okan, Garcia-Retamero, Cokely, and Maldonado (2015).
2. In all experiments, we also measured participants' numeracy (i.e., the ability to understand and manipulate different numerical expressions of probability; Lipkus, Samsa, & Rimer, 2001). We reasoned that numeracy may affect people's preferences to increase versus decrease their blood glucose, as this skill is a robust predictor of medical decisions and health outcomes (e.g., Cokely et al., 2012; Peters, 2012; Petrova et al., 2017), including glycemic control (Osborn, Cavanaugh, Wallston, & Rothman, 2010). Numeracy was assessed using the four items in the Berlin Numeracy Test (Cokely et al., 2012), together with either nine items (Experiment 1) or three items (Experiments 2 and 3) selected from the numeracy scale developed by Lipkus et al. (2001). Numeracy items were always included after the graph literacy scale. In addition, as part of the demographic questions, participants were asked to indicate whether they had a chronic disease and, in case of an affirmative response, indicate which disease. The latter questions were included as we considered that previous experience with endocrine disorders associated with glycemic control (pre-diabetes, diabetes, or thyroid disease) may also affect decisions concerning blood glucose. However, neither numeracy nor the presence of endocrine disorders were correlated with preference ratings (numeracy: $r=-.03$ in Experiments 1 and 2; presence of endocrine disorders: $r=.03$ and $r=.02$ in Experiments 1 and 2, respectively).
3. We thank Catherine Fritz for her valuable suggestions concerning this approach to analyses.
4. Considering recent recommendations for detecting inattention in online studies (Maniaci & Rogge, 2014), we computed the 5% trimmed mean completion time (17 min 46 s in Experiment 2 and 14 min 32 s in Experiment 3), and rerun our analyses excluding the participants who completed the study in less than half of this time ($n=39$ in Experiment 2 and $n=36$ in Experiment 3). All results remained unchanged, with the exception of the effect of error bars on bias in likelihood ratings in Experiment 2 (which reached conventional levels of significance, $\beta=-.08$, $t=1.99$, $p=.047$), and the interaction term between graph literacy and bars versus table in Experiment 3 (which no longer reached conventional levels of significance, $\beta=-.10$, $t=1.73$, $p=.08$). All analyses reported included the full sample. Results corresponding to the analyses with the trimmed data set for both experiments are available upon request.
5. Participants could not proceed to the next page until the Continue button had been displayed, although they could spend as much time as needed viewing each page. To avoid confusion or frustration associated with the absence of the Continue button in the initial 10 s, the following instructions were displayed at the bottom of the screen: "Click on the button that will appear below when you are ready to continue (please note that the button may NOT appear immediately, and therefore you may need to wait a few seconds until it appears)." In addition, the screen displaying the slider to assess participants' preferences included a sentence informing participants that they would be presented with information that they had already seen earlier ("Below you can view again the information presented in the last page/two pages").
6. In Experiments 2 and 3, participants also answered four questions assessing their knowledge and familiarity with blood glucose (Experiment 2) and consumption of added sugars (Experiment 3), which were presented immediately before the graph literacy scale. In Experiment 3, participants also answered two questions concerning hypothetical policy decisions, based on Stone, Gabard, Groves, and Lipkus (2015), which were included for exploratory purposes. The first question asked participants to indicate what percentage of the Centers for Disease Control and Prevention (CDC) budget they would designate for researching ways to deal

with the consumption of added sugars (vs the consumption of tobacco). The second question asked participants to assume that the CDC presently spends US$10,000 on educating the public regarding the effects of the consumption of added sugars, and asked participants to indicate their agreement with this amount. Further details are available upon request.

7. Participants in the bars and dot plot conditions were instructed to focus on the group of female between 20 and 39, and were informed that this group was represented on the right side of the graph in light blue colour. Such instructions were included to facilitate interpretation of the graphs prior to the elicitation of likelihood judgements.

References

Ali, N., & Peebles, D. (2013). The effect of gestalt laws of perceptual organization on the comprehension of three-variable bar and line graphs. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *55*, 183–203.

American Diabetes Association. (2012). Standards of medical care in diabetes—2012. *Diabetes Care*, *35*, S11–S63.

Ancker, J. S., Senathirajah, Y., Kukafka, R., & Starren, J. B. (2006). Design features of graphs in health risk communication: A systematic review. *Journal of the American Medical Informatics Association*, *13*, 608–618.

Bruine de Bruin, W., Stone, E. R., Gibson, J. M., Fischbeck, P. S., & Shoraka, M. B. (2013). The effect of communication design and recipients' numeracy on responses to UXO risk. *Journal of Risk Research*, *16*, 981–1004.

Chandler, J., & Shapiro, D. (2016). Conducting clinical research using crowdsourced convenience samples. *Annual Review of Clinical Psychology*, *12*, 53–81.

Cleveland, W. S. (1984). Graphical methods for data presentation: Full scale breaks, dot charts, and multibased logging. *American Statistician*, *38*, 270–280.

Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, *79*, 531–554.

Cokely, E. T., Feltz, A., Ghazal, S., Allan, J. N., Petrova, D., & Garcia-Retamero, R. (2018). Decision making skill: From intelligence to numeracy and expertise. In K. A. Ericsson, R. R. Hoffman, A. Kozbelt, & A. M. Williams (Eds.), *Cambridge handbook of expertise and expert performance* (2nd edition). Cambridge, UK: Cambridge University Press.

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin Numeracy Test. *Judgment and Decision Making*, *7*, 25–47.

Ervin, R. B., & Ogden, C. L. (2013). *Consumption of added sugars among U.S. adults, 2005–2010* (NCHS data brief no. 122). Hyattsville, MD: National Center for Health Statistics. Retrieved from http://www.cdc.gov/nchs/data/databriefs/db122.htm

Feldman-Stewart, D., Kocovski, N., McConnell, B. A., Brundage, M. D., & Mackillop, W. J. (2000). Perception of quantitative information for treatment decisions. *Medical Decision Making*, *20*, 228–238.

Fortin, J. M., Hirota, L. K., Bond, B. E., O'Connor, A. M., & Col, N. F. (2001). Identifying patient preferences for communicating risk estimates: A descriptive pilot study. *BMC Medical Informatics and Decision Making*, *1*, 2.

Freedman, E. G., & Shah, P. (2002). Toward a model of knowledge-based graph comprehension. In M. Hegarty, B. Meyer, & N. H. Narayanan (Eds.), *Diagrammatic representation and inference* (pp. 59–141). Berlin, Germany: Springer.

Galesic, M., & Garcia-Retamero, R. (2011). Graph literacy: A cross-cultural comparison. *Medical Decision Making*, *31*, 444–457.

Garcia-Retamero, R., & Cokely, E. T. (2013). Communicating health risks with visual aids. *Current Directions in Psychological Science*, *22*, 392–399.

Garcia-Retamero, R., & Cokely, E. T. (2017). Designing visual aids that promote risk literacy: A systematic review of health research and evidence-based design heuristics. *Human Factors*, *59*, 582–627.

Godau, C., Vogelgesang, T., & Gaschler, R. (2016). Perception of bar graphs—A biased impression? *Computers in Human Behavior*, *59*, 67–73.

Kutner, M. A., Greenberg, E., Jin, Y., & Paulsen, C. (2006). *The health literacy of America's adults: Results from the 2003 National Assessment of Adult Literacy*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Lipkus, I. M. (2007). Numeric, verbal, and visual formats of conveying health risks: Suggested best practices and future recommendations. *Medical Decision Making*, *27*, 696–713.

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making*, *21*, 37–44.

Lipska, K. J., Ross, J. S., Wang, Y., Inzucchi, S. E., Minges, K., Karter, A. J., . . . Krumholz, H. M. (2014). National trends in US hospital admissions for hyperglycemia and hypoglycemia among Medicare beneficiaries, 1999 to 2011. *JAMA Internal Medicine*, *174*, 1116–1124.

Maichle, U. (1994). Cognitive processes in understanding line graphs. In W. Schnotz & R. W. Kulhavy (Eds.), *Comprehension of graphics* (pp. 207–227). Amsterdam, The Netherlands: North Holland Elsevier Science.

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*, 61–83.

McCaffery, K. J., Dixon, A., Hayen, A., Jansen, J., Smith, S., & Simpson, J. M. (2012). The influence of graphic display format on the interpretations of quantitative risk information among adults with lower education and literacy: A randomized experimental study. *Medical Decision Making*, *32*, 532–544.

Mt-Isa, S., Hallgreen, C. E., Asiimwe, A., Downey, G., Genov, G., Hermann, R., . . . Tzoulaki, I. (2013). *Review of visualisation methods for the representation of benefit-risk assessment of medication: Stage 2 of 2*. Retrieved from http://www.imi-protect.eu/documents/ShahruletalReviewofvisualisationmethodsfortherepresentationofBRassessmentofmedicationStage2A.pdf

Newman, G. E., & Scholl, B. J. (2012). Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar bias. *Psychonomic Bulletin & Review*, *19*, 601–607.

Okan, Y., Galesic, M., & Garcia-Retamero, R. (2016). How people with low and high graph literacy process health graphs: Evidence from eye-tracking. *Journal of Behavioral Decision Making*, *29*, 271–294.

Okan, Y., Garcia-Retamero, R., Cokely, E. T., & Maldonado, A. (2015). Improving risk understanding across ability levels: Encouraging active processing with dynamic icon arrays. *Journal of Experimental Psychology: Applied*, *21*, 178–194.

Okan, Y., Garcia-Retamero, R., Galesic, M., & Cokely, E. T. (2012). When higher bars are not larger quantities: On individual differences in the use of spatial information in graph comprehension. *Spatial Cognition & Computation*, *12*, 1–25.

Okan, Y., Stone, E. R., & Bruine de Bruin, W. (2017). Designing graphs that promote both risk understanding and behavior change. *Risk Analysis*. Advance online publication. doi:10.1111/risa.12895

Osborn, C. Y., Cavanaugh, K., Wallston, K. A., & Rothman, R. L. (2010). Self-efficacy links health literacy and numeracy to glycemic control. *Journal of Health Communication*, *15*, 146–158.

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, *23*, 184–188.

Peebles, D. (2008). The effect of emergent features on judgments of quantity in configural and separable displays. *Journal of Experimental Psychology: Applied*, *14*, 85–100.

Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*, *46*, 1023–1031.

Peters, E. (2012). Beyond comprehension: The role of numeracy in judgments and decisions. *Current Directions in Psychological Science*, *21*, 31–35.

Petrova, D., Garcia-Retamero, R., Catena, A., Cokely, E., Heredia Carrasco, A., Arrebola Moreno, A., & Ramírez Hernández, J. A. (2017). Numeracy predicts risk of pre-hospital decision delay: A retrospective study of acute coronary syndrome survival. *Annals of Behavioral Medicine*, *51*, 292–306.

Pinker, S. (1990). A theory of graph comprehension. In R. Freedle (Ed.), *Artificial intelligence and the future of testing* (pp. 73–126). Hillsdale, NJ: Lawrence Erlbaum.

Shah, P., & Freedman, E. G. (2011). Bar and line graph comprehension: An interaction of top-down and bottom-up processes. *Topics in Cognitive Science*, *3*, 560–578.

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

Stone, E. R., Bruine de Bruin, W., Wilkins, A. M., Boker, E. M., & MacDonald Gibson, J. (2017). Designing graphs to communicate risks: Understanding how the choice of graphical format influences decision making. *Risk Analysis*, *37*, 612–628.

Stone, E. R., Gabard, A. R., Groves, A. E., & Lipkus, I. M. (2015). Effects of numerical versus foreground-only icon displays on understanding of risk magnitudes. *Journal of Health Communication*, *20*, 1230–1241.

Tal, A., & Wansink, B. (2016). Blinded with science: Trivial graphs and formulas increase ad persuasiveness and belief in product efficacy. *Public Understanding of Science*, *25*, 117–125.

Trevena, L. J., Zikmund-Fisher, B. J., Edwards, A., Gaissmaier, W., Galesic, M., Han, P. K. J., . . . Woloshin, S. (2013). Presenting quantitative information about decision outcomes: A risk communication primer for patient decision aid developers. *BMC Medical Informatics and Decision Making*, *13*(Suppl 2), S7.

Waters, E. A., Weinstein, N. D., Colditz, G. A., & Emmons, K. (2006). Formats for improving risk communication in medical tradeoff decisions. *Journal of Health Communication*, *11*, 167–182.

World Health Organization. (2017a). *Diabetes*. Retrieved from http://www.who.int/mediacentre/factsheets/fs312/en/

World Health Organization. (2017b). *Obesity and overweight*. Retrieved from http://www.who.int/mediacentre/factsheets/fs311/en/