ARTICLE IN PRESS

International Journal of Forecasting xxx (xxxx) xxx

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Bars, lines and points: The effect of graph format on judgmental forecasting☆

Stian Reimers [a,b,*], Nigel Harvey [b]

[a] Department of Psychology, City, University of London, UK
[b] Department of Experimental Psychology, University College London, UK

ARTICLE INFO

Keywords:
Judgmental forecasting
Time series
Format
Graph comprehension
Trend damping

ABSTRACT

Time series are often presented graphically, and forecasters often judgmentally extrapolate graphically presented data. However, graphs come in many different formats: here, we examine the effect of format when non-experts make forecasts from data presented as bar charts, line graphs, and point graphs. In four web-based experiments with over 4000 participants, we elicited judgmental forecasts for eight points that followed a trended time series containing 50 points. Forecasts were lower for bar charts relative to either line or point graphs. Factors potentially affecting these format effects were investigated: We found that the intensity of shading had no effect on forecasts and that using horizontal stepped lines led to higher forecasts than bars. We also found that participants added more noise to their forecasts for bars than for points, leading to worse performance overall. These findings suggest that format significantly influences judgmental time series forecasts.

1. Introduction

Forecasting almost always involves some degree of human judgment. This involvement can occur at different levels: Sometimes at the high level of choosing an appropriate model to provide statistical forecasts (Petropoulos, Kourentzes, Nikolopoulos, & Siemsen, 2018), frequently at the intermediate level of adjusting the output of statistical forecasting software to account for contextual knowledge (e.g., Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009), and still often at the lowest level of making forecasts using unaided judgment (Fildes & Goodwin, 2007; Fildes & Petropoulos, 2015).

Although human judgmental forecasting has an advantage over statistical forecasts by considering a wide range of background information and previous experience, it also introduces a range of well-documented biases. These include trend damping, where forecasters make predictions that are closer to the last observation than they should be given the underlying trend (e.g., Andreassen & Kraus, 1990; Harvey, 1995; Harvey & Reimers, 2013; Lawrence & Makridakis, 1989; O'Connor, Remus, & Griggs, 1997), misjudgments of serial dependence (Bolger & Harvey, 1993; Reimers & Harvey, 2011), and attempts to make forecasts look representative of the existing time series by adding noise (Harvey, 1995). (For reviews of judgmental effects in forecasting, see Bolger & Harvey, 1998; Goodwin & Wright, 1993, 1994; Harvey, 2007; Lawrence, Goodwin, O'Connor, & Önkal, 2006)

Research has also examined how some of these apparent biases may be affected by the way in which time series information is conveyed to the forecaster. This includes

☆ This research was supported in part by ESRC, UK grant RES-000-22-2007 awarded to NH and SR and by a fellowship awarded to SR by the ESRC Centre for Economic Learning and Social Evolution, UK.

* Correspondence to: Department of Psychology, City, University of London, Northampton Square, London EC1V 0HB, UK.
E-mail address: stian.reimers@city.ac.uk (S. Reimers).

https://doi.org/10.1016/j.ijforecast.2022.11.003

Please cite this article as: S. Reimers and N. Harvey, Bars, lines and points: The effect of graph format on judgmental forecasting. International Journal of Forecasting (2023), https://doi.org/10.1016/j.ijforecast.2022.11.003.

the relative merits of presenting data in graphical or tabular format (Desanctis & Jarvenpaa, 1989; Harvey & Bolger, 1996) and, recently, the precise format in which graphical information is presented (Theocharis, Smith, & Harvey, 2019). This research draws on a large body of cognitive psychology research on graphical comprehension (for reviews, see Pinker, 1990; Shah & Hoeffner, 2002; Speier, 2006), which has mapped much of the way in which presentational format can affect cognitive representation.

This paper examines the effects of three common presentation formats on unaided judgmental forecasting, specifically bar graphs, line graphs, and point graphs. Our primary argument is that if the same time series data presented in different formats leads to different forecasts, it is important to understand how and why the format has these effects. Before describing the four experiments, we review the literature – first, the cognitive psychology literature on non-forecasting graph comprehension, and then the more limited research on presentational effects in forecasting.

1.1. Format effects in graph comprehension

Graphs are used to communicate information succinctly and ergonomically in almost all domains where quantitative data are produced. Where data are complex, or the conceptualization of relationships among variables is important, graphs are substantially more effective at conveying information than numerical tables (e.g., Meyer, Shamo, & Gopher, 1999; Schonlau & Peters, 2012; Vessey & Galletta, 1991).

The perceptual, cognitive, and pragmatic processes underlying the interpretation of graphical information have been studied in several different areas (for reviews, see Pinker, 1990; Shah & Hoeffner, 2002; Speier, 2006). It is clear from many studies that the format in which graphical information is presented is not neutral – different formats affect the perception, interpretation, and recall of the data.

Some key themes on the effects of format emerge from existing research. The first concerns the difference between continuous representations like lines and discrete representations like bar charts. Shah and Hoeffner note that "a set of points may be more likely to be grouped when they are connected by a line than when they are unconnected in a bar graph" (p. 50). Zacks and Tversky (1999) reported that, even for graphs that contain just two observations, participants were more likely to describe a relationship between x and y variables as being continuous if a line graph was used than if a bar chart was used. In some cases, this even applied to dichotomous variables. For example, some participants shown line graphs with the gender on the x-axis against height described the relationship as "The more male a person is, the taller he/she is" (Zacks & Tversky, 1999, p. 148).

It is also clear that when data are presented as line graphs, participants attend more to the effects of the variable shown on the x-axis than the variable shown across separate lines. Carpenter and Shah (1998) found that in showing the effect of two binary factors (room temperature [low, high] and noise level [low, high]) on a dependent variable (exam score), the factor presented on the x-axis was seen as much more salient. In contrast, the factor presented across separate lines was rarely mentioned. However, here again, graph format mattered: where exam scores in the different conditions were shown as bars rather than lines, participants were much more likely to mention the effects of both variables (Shah and Shellhammer, cited in Shah & Hoeffner, 2002). Bar and line graphs also appear to have differing effects on identifying and interpreting trends within a time series dataset. Simcox (1984) had participants evaluate a series of bar and line graphs and indicate whether or not they were sharply increasing. He found that line graphs required a steeper gradient to be classified as sharply increasing relative to identical data presented as bar charts. In other words, trends appeared sharper when presented as bar charts than line charts. This suggests that chart format affects the identification of trends within a time series, suggesting that extrapolation of those trends in forecasting tasks might also be affected.

1.2. Format effects in judgmental forecasting

In their review of 25 years of research on judgmental forecasting, Lawrence et al. (2006) note that the way information is displayed can affect forecasts. Research has primarily been limited to a comparison between graphs and tables. For example, Harvey and Bolger (1996) found that, for un-trended series, forecasts based on graphical information were less accurate than those that presented the same data in tabular format. However, the opposite was observed for trended series: forecasts were more accurate for graphical than tabular presentations. In the latter case, the difference in error was due to participants' gross underestimation of trends presented in tables, possibly the result of anchoring to existing values.

Similarly, equivocal findings have been reported by Desanctis and Jarvenpaa (1989), who found an advantage in forecasting accuracy for data presented as graphs, or graphs and tables, over tables alone, but only after several trials' practice. Overall, the evidence appears to show that, as Lawrence et al. (2006) describe it: "Trends are better estimated from a graphical presentation, but these seem to encourage inconsistency and overforecasting when compared to tabular format" (p.498).

Perhaps one reason why the evidence concerning the relative merits of graphs and tables is less than clear-cut is that the superiority of graphs depends on the type of graph used. As we saw earlier, the processing of graphs depends on several features that can be fairly arbitrarily chosen when designing graphical displays. We examine what is probably the most salient and best-understood feature here, the format of the graph, specifically whether time series data are presented using bar graphs, line graphs, or point graphs.

From the literature on graphical perception, we would expect these different formats to be processed in subtly different ways. For example, line charts tend to emphasize the continuity of trends, whereas bar charts suggest discrete events (Zacks & Tversky, 1999).

With specific reference to forecasting, however, we might expect an extra set of phenomena to influence

trend extrapolation. One informal explanation is that points encourage the fitting of an approximate trend line, which is then extrapolated. In contrast, bars de-emphasize the underlying trend and draw attention to the noise. Furthermore, it seems clear that there is an asymmetry with bars, not present in lines and points, which emphasizes the area beneath the top of the bars over the area above them.

Tversky, Zacks, Lee, and Heiser (2000) note that although purists argue sensibly for lines to represent continuous data and bars to represent discrete data, participants in experiments "use bars and lines consistently but according to a different principle. Bars are closed forms and can be viewed as containers; they enclose one kind of thing, separating that kind of thing from other kinds of things" (p. 224), and argue that using bars for discrete relationships and lines for trends is intuitive. In our experience, professional forecasting software and most time series data use line graphs. However, newspapers, media websites, overviews of company and financial product performance, and other domains in which time series are regularly presented use an inconsistent mix of line and bar graphs. Indeed, it is not always clear when a variable should be treated as continuous and therefore shown as a line or discrete. For example, annual sales figures for several years represent a practically continuous variable but one that has been grouped into discrete time slices. On the other hand, variables that have a constantly fluctuating instantaneous value, such as exchange rates or stock prices, are almost always shown as line graphs (even though the line graphs rarely show the continuous data but instead depict samples taken at regular intervals). In this case, it would be inappropriate to represent the data using bar charts, which would give the impression of discrete events.

One phenomenon that potentially explains why forecasts may be different for bar graphs versus other graphs is the within-the-bar effect. Newman and Scholl (2012) presented participants with graphs containing a single bar showing the mean of a sample. They asked them to rate the likelihood that a point a certain distance from the top of the bar would come from the distribution depicted by the bar. They found that points above the top of the bar were rated as less likely to come from the distribution than those an equal distance below the top (and hence in the shaded area within the bar). The authors controlled for various potential explanations, such as biases towards higher elevations, and argued that the asymmetry occurred because the way the bar was perceived led participants to erroneously judge a point within the enclosed, shaded bar area as being more likely to come from the distribution. We know of a single study that has examined something similar to the within-the-bar effect in trended data. Correll and Heer (2017) asked participants to adjust a best-fit function to a series of 100 data points presented either as a point graph, a line graph, or a line graph with the area below the line filled (to mimic the asymmetry of the within-the-bar effect). As well as finding that increased noise in the series led to worse performance, they found that estimated best-fit lines for area charts were lower than those for line charts. Although this study did not use a forecasting task, it suggests that forecasting may also be affected by format similarly.

The only study we know of that directly tested the effects of format on forecasting was that of Theocharis et al. (2019), who examined forecasting to real time series of annual hurricane frequencies. Across a series of different forecasting tasks, they found that forecasts using line graphs showed more serial dependence than those using point graphs, with forecasts closer to the last observation for lines than points. They argued that connecting point observations with lines created a sense of interconnection, which increased the effect of the last observations on the forecast.

1.3. Rationale

A large body of work in the judgmental forecasting literature has pointed to unaided forecasts being systematically biased in several increasingly well-documented ways. A separate body of work on graph comprehension suggests that the format of graphs can strongly influence how they are perceived. However, there is very little work on the extent to which graph format affects judgmental forecasting. We aim to fill that gap here with a series of experiments in which participants make forecasts to identical time series presented as bar graphs, line graphs, and point graphs. We then examine potential mechanisms by which these differences may occur by separately manipulating factors like format, shading, and gradient of connecting lines.

We predict that:

H1: (a) Forecasts using bars will be lower than those using lines or points, and (b) the effect of the format will increase with a more distant time horizon.

H2: Participants will add more noise to forecasts using bars than those using lines or points.

We test these predictions in Experiments 1a and 1b. Experiments 2 and 3 investigate the underlying mechanisms behind the results obtained in Experiment 1.

2. Experiment 1a

Here we examine whether judgmental forecasts of naïve forecasters are affected by the way format in which time-series graphs are displayed. We use five different functions and two levels of normally distributed noise in a between-participants design. Although we use multiple functions, this is primarily to ensure that results are not specific to a single function type and to give an initial sense of how variable any format effects might be across different types of time series. We make no theoretical predictions about how function type might interact with the other independent variables. Although we report interactions involving function, we focus on the theoretically motivated analyses of format effects.

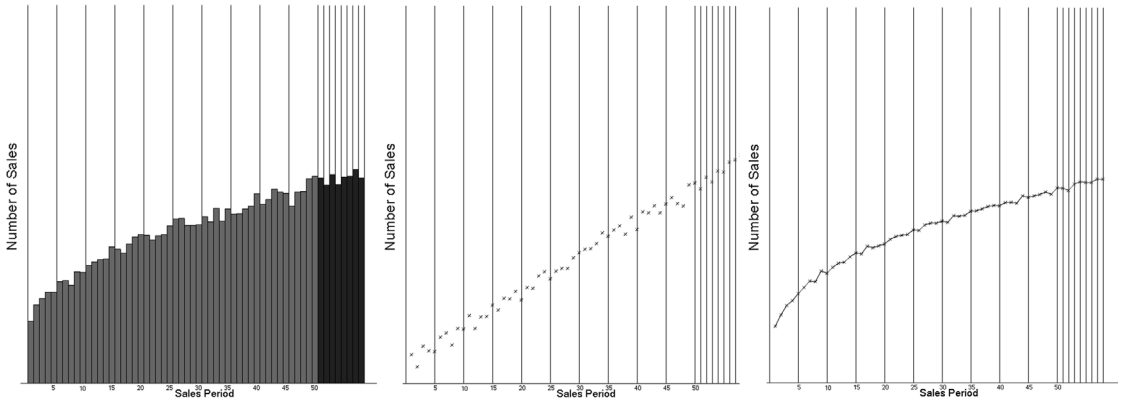

Fig. 1. The three formats used in Experiments 1a and 1b. The left panel shows a decelerating, high-noise function in Bars format; The middle panel shows a steep linear high-noise function in Point format; The right panel shows a decelerating, low-noise function in Lines format.

2.1. Method

2.1.1. Participants

A total of 1069 participants were recruited using the ipoints scheme (www.ipoints.co.uk), an internet reward scheme that allows members to collect points to exchange for things like CDs, electronic equipment, and shopping vouchers. Participants were paid 40 ipoints (with a trade-in value of around 15p, $0.25) for completing the two-minute experiment. The 50 participants whose forecasts were closest to the noiseless trend line in MAE were awarded a further 75 ipoints, and the participant whose forecast was closest received a bonus of 1000 points

Materials. The trend functions used were a subset of those used in previous research (e.g., Harvey & Reimers, 2013). These trends were designed so that the observation of the time series was in the same position in the middle of the *y*-axis on the graph, which had a height of 700 pixels.[1] They were as follows:

Decelerating: $y = 50 + 300\ (x/\ 50)^{0.4}$
Accelerating: $y = 50 + 300\ (x/\ 50)^{1.5}$
Shallow Linear $y = 200 + 300\ (0.5x/\ 50)$
Medium Linear $y = 125 + 300\ (0.75x/\ 50)$
Steep Linear $y = 50 + 300\ (x/\ 50)$

Time series were presented in three formats: bar graphs, line graphs, and point graphs. Examples of these formats are shown in Fig. 1 (a full set of stimulus types is given in Supplementary Materials S3). As described below, random noise was added to each function for each participant.

2.1.2. Design

The experiment was conducted between participants to avoid trial-to-trial carryover effects (see Harvey & Reimers, 2013) and to avoid making the experimental manipulations obvious to participants. There were three between-participant factors: function (5 levels: Shallow Linear, Medium Linear, Steep Linear, Negatively Accelerated, Positively Accelerated), Gaussian noise (2 levels: Low [M = 0, SD = 3] and High [M = 0, SD = 10]), and format (3 levels: Bars, Lines, Points). This gave a total of 30 cells. Participants were randomly allocated to one of the 30 cells.

2.1.3. Procedure

The experiment was coded in Flash (Reimers & Stewart, 2007, 2015) and run online. Participants received an email inviting them to take part by following a link to the URL at which the experiment was hosted. At the start of the experiment, participants were given the following instructions:

> You'll take the role of an advisor to a company. You'll see the company's sales figures for the past 50 sales periods. Your job is to make your forecast for the next 8 sales periods as accurately as you can. You do this by clicking on the lines just beyond the existing sales figures. By using your judgment based on the existing trend, you should be able to make a fairly accurate prediction. There are no tricks here - the trend you see is based on real trends seen in business forecasting.
>
> As an incentive, there are bonuses for being in the top 50 respondents who make the most accurate forecasts, and an overall 1000-ipoint prize for the most accurate response overall.

Next, participants gave their email addresses (to allow payment) along with their age, gender and education. They then completed the forecasting task, being shown a time series of 50 observations and generating the next eight observations by clicking to the right of the existing series. Participants clicked in the vertical space following the 50 points, and a bar, line, or point appeared where they had clicked. The bar's position, line, or point could be changed by clicking again in a different location. For participants in the line condition, each time they added a forecast, the software running the experiment joined the forecast to forecasts on either side if any were present with a line. The eight points to be forecasted could be entered in any order. Once participants had made all eight forecasts and were happy with them, they pressed a 'submit' button to proceed.

[1] In Experiments 1 and 2 there was an inadvertent vertical shift of 10 pixels to time series positions. This affected all conditions, so is unlikely to have had any effect on the pattern of results.

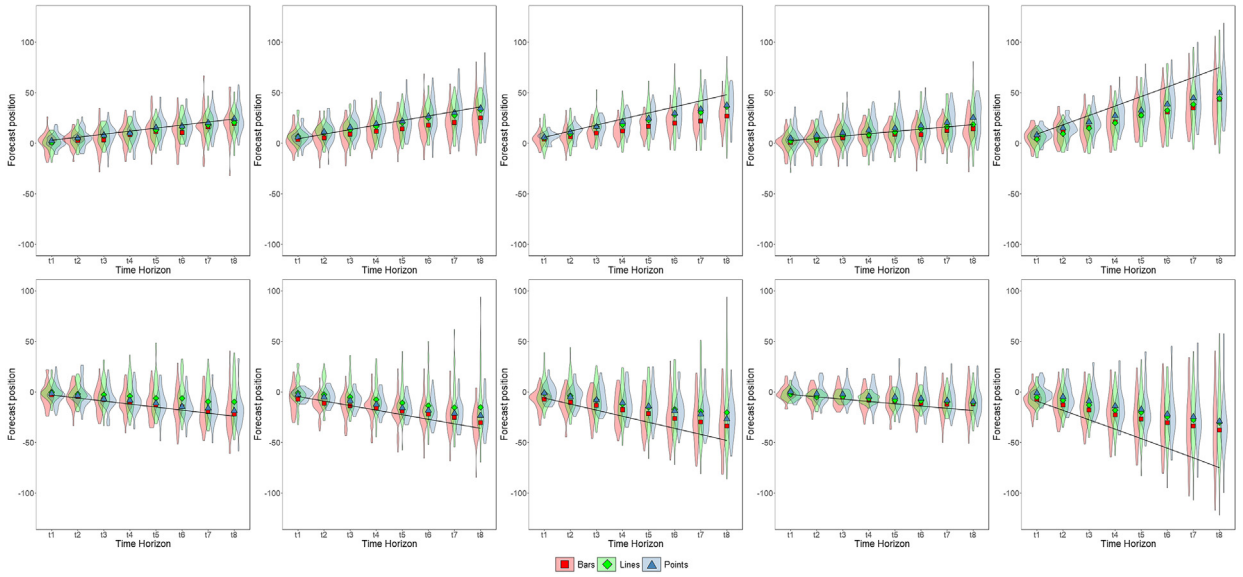

Fig. 2. Mean forecasts for the five trend functions, collapsing across noise, in Experiments 1a and 1b (Left to Right: Shallow Linear, Medium Linear, Steep Linear, Decelerating, Accelerating), for upward (Top Panel) and downward (Bottom Panel) trends. Point means are given (Bars: Red, Lines: Green, Points: Blue), and violin plots show the distribution of responses from left to right for bars, lines, and points. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2.2. Results

Overall, 33% of participants were female, and 67% were male; the median age was 45 years old (IQR: 18 years). As data were relatively noisy, we removed outlying forecasts in the same way for all experiments reported here, a method identical to that previously used (Harvey & Reimers, 2013). This was an iterative procedure in which any of the eight forecasts that lay more than two interquartile ranges below the lower quartile or above the upper quartile for their condition led to the removal of that participant from the final analysis. After outlier removal, the procedure was repeated until no further participants were removed. This process aimed to remove participants who were not attempting to make accurate forecasts, for example, by clicking on regions of the screen nearest the 'Next' button, but retain participants whose attempts were genuine, even if they were internally noisy or inaccurate. The general pattern of results was not particularly sensitive to outlier removal: If outliers were not removed and all data were included in the analysis, the effects became substantially weaker, but the interaction between graph format and time horizon, and general finding of lower forecasts to bars remained. See Supplementary Materials S1 for a complete comparison of different outlier removal strategies. See Supplementary Materials S2 for tables of means for each cell across all experiments reported here. Raw data and analysis scripts, which allow comparison between trimmed and non-trimmed analyses, are archived at: https://osf.io/ztx4c/?view_only=dc21a2ff2a5f40f789d4214e1f5256b5.

After iteratively removing outliers, data from 856 participants remained in the analysis. The pattern of raw forecasts across the five functions and three formats can be seen in the top row of Fig. 2. The number of participants in each of the 30 cells varied from 15 to 37 (median = 28.5). First, we subjected the data to the mixed analysis of variance (ANOVA) with three between-participant variables (Function, Noise, Format) and one within-participant variable (Time Horizon). Here, and in later experiments, where sphericity was violated, we used the Greenhouse–Geisser method to correct the degrees of freedom, hence the non-integer degrees of freedom for some analyses. We report generalized eta-squared as a measure of effect size here and throughout (Olejnik & Algina, 2003).

The results of the analysis are summarised in Table 1. An interaction between format and time horizon arose because the effects of format increased with time horizon. Further analysis showed a significant effect of format at all levels of the time horizon ($Fs = 6.83 – 20.4$, $\eta^2_G = .016 – .047$), and at all levels, the same rank ordering (bars below lines below points) was observed. Therefore, the main effect of format was unsurprisingly significant (Table 1). Tukey tests showed a significant difference between all three formats (ps = .002 – <.001) with forecasts using bars lowest, lines next, and points highest. Other effects reported in Table 1 arose because forecasts and how they changed with time horizon varied, unsuprisingly, with function type and noise level.

Next, we wanted to examine whether participants added more noise to their forecasts in some conditions than others. This analysis ignores the data's actual trend and focuses on the amount of variability in participants' responses. To calculate this, we took each participant's forecasts for the eight levels of time horizon. We fitted a regression to them with linear and quadratic components, using the root mean square error (RMSE) to measure the amount of noise added to forecasts.[2]

[2] In much judgmental forecasting research, RMSE is used to compare participant forecasts to the underlying trend line, as a normative

Table 1
Summary of Effects and interactions in Experiment 1a.

Variable	Degrees of freedom	F ratio	p	η^2_G
Function	**4, 826**	**55.4**	**<.001**	**.16**
Format	**2, 826**	**20.4**	**<.001**	**.03**
Noise	1, 826	0.40	.527	<.001
Function x Format	8, 826	1.36	.213	.01
Function x Noise	**4, 826**	**4.49**	**.001**	**.02**
Format x Noise	2, 826	1.91	.149	.003
Function x Format x Noise	8, 826	0.84	.568	.006
Time Horizon	**3.2, 2602**	**1200**	**<.001**	**.30**
Function x Time Horizon	**12.6, 2602**	**31.1**	**<.001**	**.04**
Format x Time Horizon	**6.3, 2602**	**6.43**	**<.001**	**.004**
Noise x Time Horizon	3.2, 2602	0.99	.397	<.001
Function x Format x Time Horizon	25.2, 2602	0.62	.93	.002
Function x Noise x Time Horizon	**12.6, 2602**	**4.56**	**<.001**	**.006**
Format x Noise x Time Horizon	6.3, 2602	1.44	.194	.001
Function x Format x Noise x Time Horizon	25.2, 2602	0.95	.536	.003

Table 2
Summary of effects and interactions on RMSE to individual forecasts across the time horizon.

Variable	Degrees of freedom	F ratio	p	η^2_G
Format	**2, 826**	**5.48**	**.004**	**.013**
Function	4, 826	0.96	.427	.005
Noise	**1, 826**	**493**	**<.001**	**.374**
Format x Function	**8, 826**	**2.05**	**.038**	**.019**
Format x Noise	**2, 826**	**6.69**	**.001**	**.016**
Function x Noise	4, 826	0.30	.880	.001
Format x Function x Noise	8, 826	1.76	.082	.017

We used RMSE as our dependent variable in a 3-factor ANOVA (Format, Noise, Function Type), with results summarized in Table 2. There was an interaction between noise and format, which appears to be the result of the difference between the noisiness of forecasts in low and high noise conditions being larger for bars than lines or points. Looking at the effects of format at each level of noise, in the low noise condition there was no evidence of a format effect, $F(2, 439) = 2.09$, $p = .125$, $\eta^2_G = .009$, but for the high noise condition it was very clear, $F(2, 387) = 6.31$, $p = .002$, $\eta^2_G = .032$. There was also an interaction between format and function type. Still, there was no significant effect of format at any individual level of function type, presumably because of the reduced power of examining each level in isolation.

Post-hoc pairwise comparisons using a Tukey test showed that RMSE was significantly higher with the Bars format than the Points format ($p = .003$), suggesting that participants added more noise to their forecasts in the Bars condition. There was no difference between Bars and Lines ($p = .11$) and Lines and Points ($p = .34$)

These results suggest that when people make forecasts using bar graphs, they make lower predictions than those made to the line and point graphs with the same data. There are two potential explanations: an absolute effect – participants just forecast lower with bars whatever trend they are asked to extrapolate; or a damping effect – participants damp trends more when using bar charts, drawing their forecasts towards the horizontal. As Experiment 1a used only trends with positive gradients, increased damping would lead to overall lower forecasts. To choose between these two accounts, Experiment 1b replicates Experiment 1a using mirror-image downward trends. With downward trends, increased damping for bars would lead to higher rather than lower forecasts.

3. Experiment 1b

3.1. Method

Experiment 1b used a very similar design to Experiment 1, except that functions were flipped horizontally, meaning that all functions were now downward trends.

3.1.1. Participants

A total of 1063 participants, recruited as in Experiment 1a from the ipoints panel, completed the experiment. No participants who had completed Experiment 1a were allowed to take part in Experiment 1b.

3.2. Results

Overall, 65% of participants were female, and 34% were male; the median age was 40 years old (IQR: 20 years). After removing outliers using the same process as in

measure of accuracy. Note that the measure here is different: It is a fit to participants' own forecasts, and does not attempt to compare with a normative baseline. As such it is a relatively pure measure of variability across participants' eight point forecasts – at the extreme, eight forecasts on a straight line or quadratic curve would have zero RMSE.

Table 3
Summary of Effects and interactions in Experiment 1b.

Variable	Degrees of freedom	F ratio	p	η^2_G
Function	**4, 753**	**13.3**	**<.001**	**.048**
Format	**2, 753**	**12.9**	**<.001**	**.024**
Noise	**1, 753**	**9.06**	**.003**	**.009**
Function x Format	8, 753	1.30	.242	.010
Function x Noise	4, 753	0.47	.754	.002
Format x Noise	2, 753	1.91	.148	.004
Function x Format x Noise	**8, 753**	**2.76**	**.005**	**.021**
Time Horizon	**2.2, 1663**	**334**	**<.001**	**.111**
Function x Time Horizon	**8.8, 1663**	**7.08**	**<.001**	**.010**
Format x Time Horizon	4.4, 1663	1.93	.095	.001
Noise x Time Horizon	2.2, 1663	1.57	.205	<.001
Function x Format x Time Horizon	17.7, 1663	0.78	.729	.002
Function x Noise x Time Horizon	8.8, 1663	1.48	.153	.002
Format x Noise x Time Horizon	4.4, 1663	0.47	.774	<.001
Function x Format x Noise x Time Horizon	**17.7, 1663**	**1.75**	**.027**	**.005**

Table 4
Summary of effects and interactions on RMSE to individual forecasts across the time horizon.

Variable	Degrees of freedom	F ratio	p	η^2_G
Format	**2, 753**	**8.41**	**<.001**	**.022**
Function	4, 753	0.61	.655	.003
Noise	**1, 753**	**344**	**≤.001**	**.314**
Format x Function	8, 753	1.40	.192	.015
Format x Noise	**2, 753**	**4.74**	**.009**	**.012**
Function x Noise	**4, 753**	**2.45**	**.045**	**.013**
Format x Function x Noise	8, 753	1.26	.262	.013

Experiment 1a, data from 783 participants entered the analysis. As before, we subjected the data to a mixed ANOVA with three between-participant variables (Function, Noise, Format) and one within-participant variable (Time Horizon).

The results of the ANOVA are given in Table 3. This time there was no interaction between Format and Time Horizon, and the effect of format can be seen in its main effect. Tukey posthoc tests showed that forecasts with bars were significantly lower than those with points or lines (*ps* <.001), but there was no significant difference between lines and points. Other effects reported in Table 3 arose because forecasts and how they changed with time horizon varied with function type and noise level.

Next, we looked at RMSE to a quadratic fit across the eight levels of the time horizon to examine the amount of noise participants added to their forecasts (Table 4). There were main effects of Format and Noise. As in the previous experiment, post hoc pairwise comparisons using a Tukey test showed that RMSE was significantly higher with the Bars format than the Points format ($p < .001$), suggesting that participants added more noise to their forecasts in the Bars condition. There was an interaction between Format and Noise. As in Experiment 1a, in the low noise condition, there was no evidence of a Format effect, $F(2, 409) = 0.71$, $p = .491$, $\eta^2_G = .003$, but for the high noise condition, it was evident, $F(2, 344) = 7.24$, $p < .001$, $\eta^2_G = .040$.[3]

[3] We do not report analyses around adding noise for subsequent experiments that used only lines and bars, as they are tangential to the paper's aims, but the general pattern of results was of minimal differences in RMSE between line and bar formats, and highly significant effects of noise where two levels of noise were used.

As with Experiment 1a, the main finding from these analyses is that forecasts were lower with bars than with points or lines, implying that forecasts are generally lower with bars rather than solely damped more. It is, however, clear from comparing participants' judgments to the trend lines in Fig. 2 that in all conditions, trend damping occurs.

3.3. Discussion

Experiments 1a and 1b showed that participants made lower predictions when forecasting using bar graphs than either line or point graphs, supporting H1a. This tendency was seen for both upward (Experiment 1a) and downward (Experiment 1b) trends, so it cannot be accounted for by assuming that damping increases when people forecast using bars. In Experiment 1a, there was an interaction between Format and Time Horizon, with forecasts using bars becoming increasingly lower with increasing time horizon (see Fig. 2); however, no similar interaction was found for downward trends in Experiment 1b. This pattern gives partial support for H1b. As well as making lower forecasts with bars, participants added more noise to their forecasts with bars, as captured by RMSE to a best-fit regression function through participants' forecasts, supporting H2. We also found clear evidence of trend damping for both upward and downward trends (see Fig. 2).

Whether these findings make bar charts unsuited for forecasting is an open question. It is inappropriate to add noise to forecasts – error is minimized by extrapolating the underlying trend rather than attempting to match the noisy appearance of the time series. Nevertheless, it is established that people do this (Harvey, 1995). One potential interpretation is that bar charts appear to draw attention to the variability around the underlying

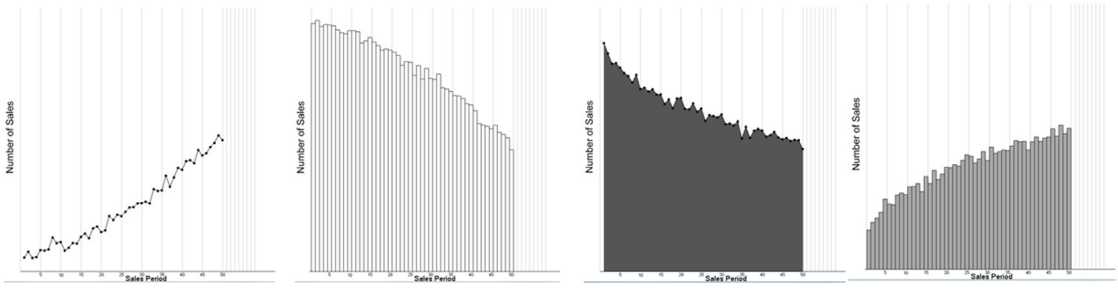

Fig. 3. Examples of shading manipulation used in Experiments 2a and 2b.

trend, thereby increasing people's tendency to add the noise they perceive to be present in the series to their forecasts. If this were the case, their forecasts would be less accurate with bars. On the other hand, the general effect of finding lower forecasts using bars may or may not be desirable, depending on the context. This is illustrated by the fact that for upward trends, forecasts using bars showed more directional error, as measured by mean deviation from the extrapolated trend line, compared to lines and points. However, forecasts with bars showed *less* directional error for downward trends than lines and points. This is readily understandable by considering the effects of trend damping, which was observed in all format conditions. As discussed above and shown in Fig. 2, forecasters engage in trend damping: they act as if they perceive the trend in the series to be less than it is, and, as a result, their forecasts are drawn towards the horizontal. For upward trends, this means that forecasters underestimate future outcomes and that using bar charts leads to an even greater underestimation the future trends. Conversely, for downward trends, trend damping means that forecasts *over*estimate future outcomes. In these latter circumstances, using bar charts corrects some of this damping by lowering forecasts. Bar charts might also be useful for improving forecasts where participants are prone to optimism biases (Harvey and Reimers, 2013, Figure 10). For example, Fildes et al. (2009) analyzed 60,000 forecast triples (initial statistical forecasts, judgmentally adjusted final forecasts, outcomes) obtained from four supply-chain companies. They found that "positive adjustments, which involved adjusting the forecast upwards, were much less likely to improve accuracy than negative adjustments. They were also made in the wrong direction more frequently, suggesting a general bias towards optimism" (p. 3). If these inappropriate upward adjustments reflect optimism, it is possible that presenting demand planners with sales series and statistical forecasts in a bar format would reduce this bias and thereby decrease damaging upward adjustments.

Having shown in two experiments that forecasts are lower when forecasting with bars, we now attempt to investigate the causes of these effects and the factors that modulate them. We address these issues in two further experiments. One potential explanation is the asymmetry in bar graphs – the area below the observation is shaded and marked as important, and attention is drawn to that area. In line graphs, the region below the line is the same color as above. As such, bar charts tend to draw attention to the region of the chart that lies below the observed data, and this may bias people toward responding in that region. For example, if participants' attention was drawn to the center of the bar or stochastically to a random position within the bar, that would represent a low anchor (Tversky & Kahneman, 1974). Through insufficient adjustment or localized activation (Chapman & Johnson, 1999), this would be expected to lead to lower forecasts.

This precise explanation is speculative. However, it could be that the more salient the area under a chart is, the more people will tend to make lower forecasts. Experiment 2 looked at both line and bar graphs, manipulating the shading beneath both graphs to examine the extent to which the emphasis and asymmetry brought by shading are responsible for drawing predictions downwards. (Clearly, it is not possible to do this for scatterplots.) We use three conditions: One in which the area beneath the trend – both for lines and bars – was unfilled, one in which it was filled with a light color, and one in which it was filled with a dark color. Our hypotheses here are:

H3: There will again be a main effect of format, with forecasts using bars being lower than forecasts using lines, and this effect will increase across the time horizon.

H4: There will be an effect of shading, with more darkly shaded areas under the trend leading to lower forecasts than lighter or unshaded versions of the same graphs.

4. Experiment 2a

In Experiment 2, we cut down the between-subject variables' levels to maximize per-participant power. This was done by retaining just the high level of noise, in which greater format effects were observed in Experiments 1a and 1b, and three function types (Accelerating, Medium Linear, and Decelerating) to ensure variety in both gradient and curvature.

4.1. Method

4.1.1. Participants

Participants were recruited using the same method as in Experiment 1 from the same participant pool. A total of 576 participants completed the experiment.

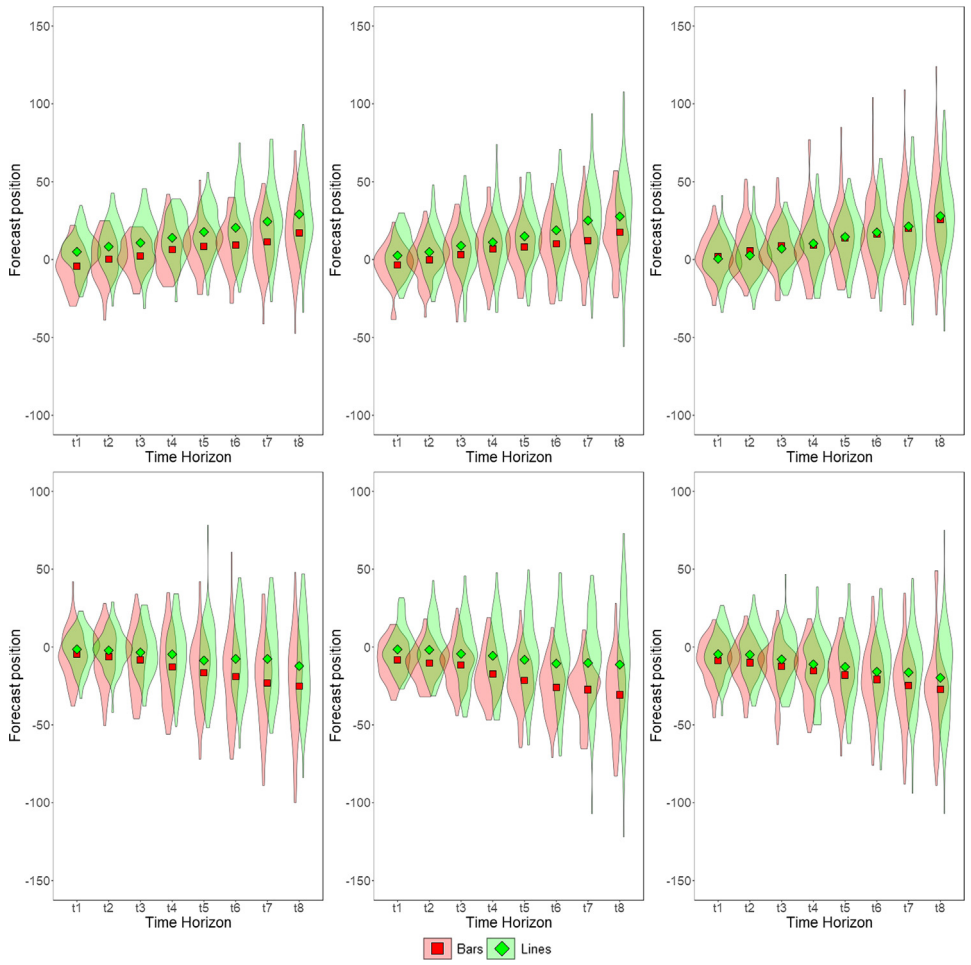

Fig. 4. Mean forecasts for Experiment 2a (Top Panels) and Experiment 2b (Bottom Panels) for No Shading (Left Panels), Medium Shading (Middle Panels), and Dark Shading (Right Panels) collapsed across noise and function.

4.1.2. Design and procedure

As in Experiment 1, Experiment 2a used a completely between-subjects design. Participants made eight forecasts for a single trend. The factorial between-participants design comprised the following variables: Format (Lines, Bars), beneath-the-trend shading (None, Light, Dark), and Function Type (Accelerating, Linear, Decelerating). As before, participants saw a noisy trend of 50 observations and then made a forecast for the next eight observations. Fig. 3 shows the different shading conditions in each format.

4.2. Results

Data from six participants who had completed one of the other studies reported here were removed, leaving 570 participants, of whom 53% were female and 46% were male. The median age was 39 years old (IQR: 21 years). After outlier removal, data from 456 participants entered the analysis. Mean forecasts are shown in the top row of Fig. 4; a summary of the ANOVA is given in Table 5.

As in Experiment 1, there was an interaction between format and time horizon. Examining the effects of format at each level of the time horizon showed a significant format effect for every level of time horizon (Fs = 5.86 – 21.7, ps = .02 – <.001, η^2_G = 013 – .047). Unsurprisingly there was, therefore, also a main effect of format.

There was also an interaction between format and shading. Examining each level of shading separately, we found an apparent effect of format for no shading, F (1, 135) = 17.9, p < .001, η^2_G = .08, and light shading conditions, F (1, 152) = 9.91, p = .002, η^2_G = .05, but no effect in the dark shading condition, F (1, 151) = 0.02, p < .89, η^2_G < .001.

Overall, this suggests that shading has minimal effect on the overall elevation of forecasts but may have very specific effects, most notably potentially having dark shading reduce the difference in forecasts to Bars and Lines.

5. Experiment 2b

Experiment 2b replicates Experiment 2a using downward trends.

Table 5
Summary of Effects and interactions in Experiment 2a.

Variable	Degrees of freedom	F ratio	p	η^2_G
Function	**2, 438**	**5.31**	**.005**	**.016**
Format	**1, 438**	**19.0**	**<.001**	**.029**
Shading	2, 438	1.42	.24	.004
Function x Format	2, 438	0.54	.59	.002
Function x Shading	**4, 438**	**4.26**	**.002**	**.026**
Format x Shading	**2, 438**	**4.52**	**.01**	**.014**
Function x Format x Shading	4, 438	1.03	.39	.006
Time Horizon	**3.6, 1565**	**245**	**<.001**	**.151**
Function x Time Horizon	**7.1, 1565**	**5.78**	**<.001**	**.008**
Format x Time Horizon	**3.6, 1565**	**4.91**	**.001**	**.004**
Shading x Time Horizon	7.1, 1565	1.09	.37	.002
Function x Format x Time Horizon	7.1, 1565	1.78	.09	.003
Function x Shading x Time Horizon	**14.3, 1565**	**1.89**	**.02**	**.005**
Format x Shading x Time Horizon	7.1, 1565	0.40	.91	.001
Function x Format x Shading x Time Horizon	14.3, 1565	0.76	.71	.002

Table 6
Summary of Effects and interactions in Experiment 2b.

Variable	Degrees of freedom	F ratio	p	η^2_G
Function	2, 396	1.92	.15	.007
Format	**1, 396**	**20.9**	**<.001**	**.036**
Shading	2, 396	1.68	.19	.006
Function x Format	2, 396	0.02	.98	<.001
Function x Shading	4, 396	0.40	.81	.003
Format x Shading	2, 396	1.10	.33	.004
Function x Format x Shading	4, 396	0.94	.44	.007
Time Horizon	**2.7, 1079**	**76.9**	**<.001**	**.053**
Function x Time Horizon	5.4, 1079	2.05	.06	.003
Format x Time Horizon	**2.7, 1079**	**7.65**	**<.001**	**.005**
Shading x Time Horizon	5.4, 1079	0.38	.88	.001
Function x Format x Time Horizon	5.4, 1079	0.44	.83	.001
Function x Shading x Time Horizon	10.9, 1079	0.98	.46	.003
Format x Shading x Time Horizon	5.4, 1079	0.93	.46	.001
Function x Format x Shading x Time Horizon	10.9, 1079	1.02	.42	.003

5.1. Method

5.1.1. Participants

In total, 581 submissions were made. Data from 12 participants who had completed one of the other studies reported here were excluded, leaving 569 participants' data, of whom 54% were female, 45% were male, and 1% did not report their gender; the median age was 40 years old (IQR: 20 years). These data were subjected to the outlier removal procedure, leaving 414 in the final analysis.

5.1.2. Design and procedure

The design and procedure were identical to Experiment 2a.

5.2. Results

Results were similar to Experiment 2a and are shown in the bottom row of Fig. 4. As before, a mixed ANOVA contained the following between-participants variables: Format, Shading, and Function Type, along with the repeated measure of Time Horizon. As before, the dependent measure was the absolute value of forecasts made by participants. Results are summarised in Table 6.

As in Experiment 2a, there was an interaction between Format and Time Horizon. Examining the effects of format at each time horizon separately, we found significant format effects at every level (Fs = 9.20 – 26.7, ps = .002 – <.001, η^2_G = .023 – .063). It is, therefore, unsurprising that overall there was also a main effect of format.

However, there was no main effect of shading nor interaction between Shading and Time Horizon.

5.3. Discussion

The results of this experiment are clear: We replicated the main findings of Experiment 1, that forecasts were lower using bars than using lines, and that for both upward and downward trends, forecasts diverged with increasing time horizon. This supports both parts of hypothesis H3. However, we found no main effect of shading, allowing us to reject H4, although we found some interactions involving shading.

This implies that bars are not forecast lower than lines because the area below bars tends to be shaded, whereas the area below lines does not. Although we found the same consistent lower forecasting with bars as in the previous experiment, this was unaffected by whether the bars were shaded or whether the area below the lines was shaded. Therefore, we can largely rule out the asymmetry in coloring below and above the time series as the cause

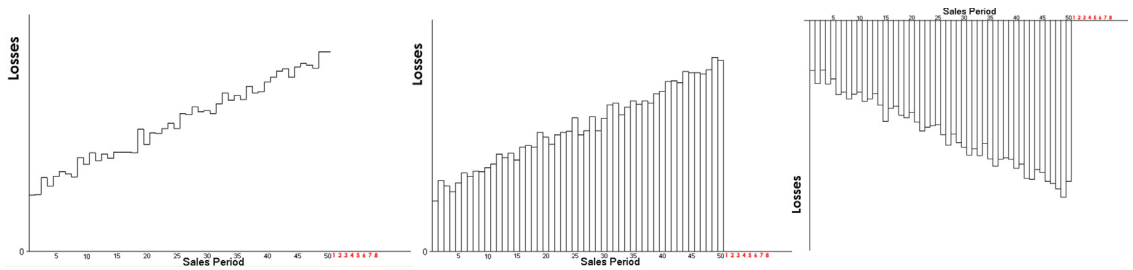

Fig. 5. Examples of stimuli used in Experiments 3 and 4: Stepped lines (Left Panel), rising bars (Middle Panel), and hanging bars (Right Panel).

of lower forecasts for bars. This is somewhat different from the findings of Correll and Heer (2017), who found that shading the area beneath a line graph led to lower estimates of the best fit line through the data. Of course, the two studies differed in several ways – Correll and Heer were not using a forecasting task – but it suggests that there may be circumstances in which shading does affect judgment. We certainly would not rule out the potential effects of shading on forecasting; however, it is clear that it is not shading that underlies the difference between bars and lines in the types of forecast participants made in our experiments.

Having established that shading does not drive differences we observed in forecasting to line and bar graphs; we turn to other differences between the two formats. Of course, even in unshaded bar graphs, there is still an asymmetry – the region under the bar has the rising lines that form the bar's perimeter, whereas the region above the bar contains nothing. If this drives the difference between bar and line graphs forecasts, it would fit more closely with the gestalt notion of enclosure (Wertheimer, 1923). On the other hand, it may be that differences in the information presented in a line graph lead to different forecasts – line graphs represent the distribution of gradients between two points. In contrast, bar graphs emphasize the difference in absolute values between two points.

In this next experiment, we attempted to create a line graph that was as similar as possible to bar graphs but without the asymmetry of the bars. To do this, we used a stepped line chart. This is identical to the type of bar chart we have used before, except that it does not have vertical lines to make the bars (see Fig. 5, Left Panel). Using this approach, we could also use the same procedure for drawing the bar and line charts, leaving out the long vertical bar lines in the stepped line condition but otherwise drawing the time series in the same way. This means that the same information is presented for both bars and stepped lines, except for the attentional and conceptual asymmetry of having vertical lines giving the impression of bars. For both bars and stepped lines, we removed the vertical reference lines present in previous experiments (compare Figs. 1 & 3 with Fig. 5). This was to avoid creating implicit bars in the stepped line condition, undermining differences between the conditions, which here only differ in the presence or absence of vertical lines. We hypothesize:

H5: Forecasts using stepped lines will be higher than those using bars, and they will diverge with increasing time horizon

6. Experiment 3a

The design of Experiments 3a and 3b was very similar to that of Experiments 1a and 1b, except that there were two format conditions: Bars and Stepped Lines. We used two levels of noise as before and the same three functions as in Experiment 2. In Experiment 3a all the trends were upward, and in Experiment 3b, all the trends were downward.

6.1. Method

6.1.1. Participants

Participants were recruited using the same method as in Experiment 1, from the same participant pool. A total of 841 participants completed the experiment.

6.1.2. Design and procedure

As in Experiment 1a, Experiment 3a used a fully between-participants design for the stimulus types: it comprised the following variables: Format (Stepped Lines, Bars), Noise (Low, High), and Function Type (Accelerating, Linear, Decelerating). As before, participants saw a noisy trend of 50 observations and then made a forecast for the next eight observations.

6.2. Results

Data from 85 participants who had completed one of the other studies reported here were removed, leaving 758 participants; 54% were female, 45% were male, and 1% did not give their gender. The median age was 42.5 years old (IQR: 25 years). After removing outliers, 567 participants remained in the analysis. As before, we subjected the data to mixed ANOVA, which this time had the three between-participants variables of Function, Noise, and Format, and one within-participant variable (Time Horizon). A summary of the analysis can be seen in Table 7, and the general pattern of format effects can be seen in Fig. 6, top left panel.

As in earlier studies, there was an interaction between Format and Time Horizon. Examining the effects of format at each level of time horizon separately, we found significant format effects at every level (Fs = 14.9 – 44.6, ps <.001, η^2_G = .026 – .074). It is, therefore, unsurprising that overall there was also a main effect of format. There were also various effects and interactions involving Function and Noise, suggesting that forecasts are affected by conjunctions of function type and noise in complex ways.

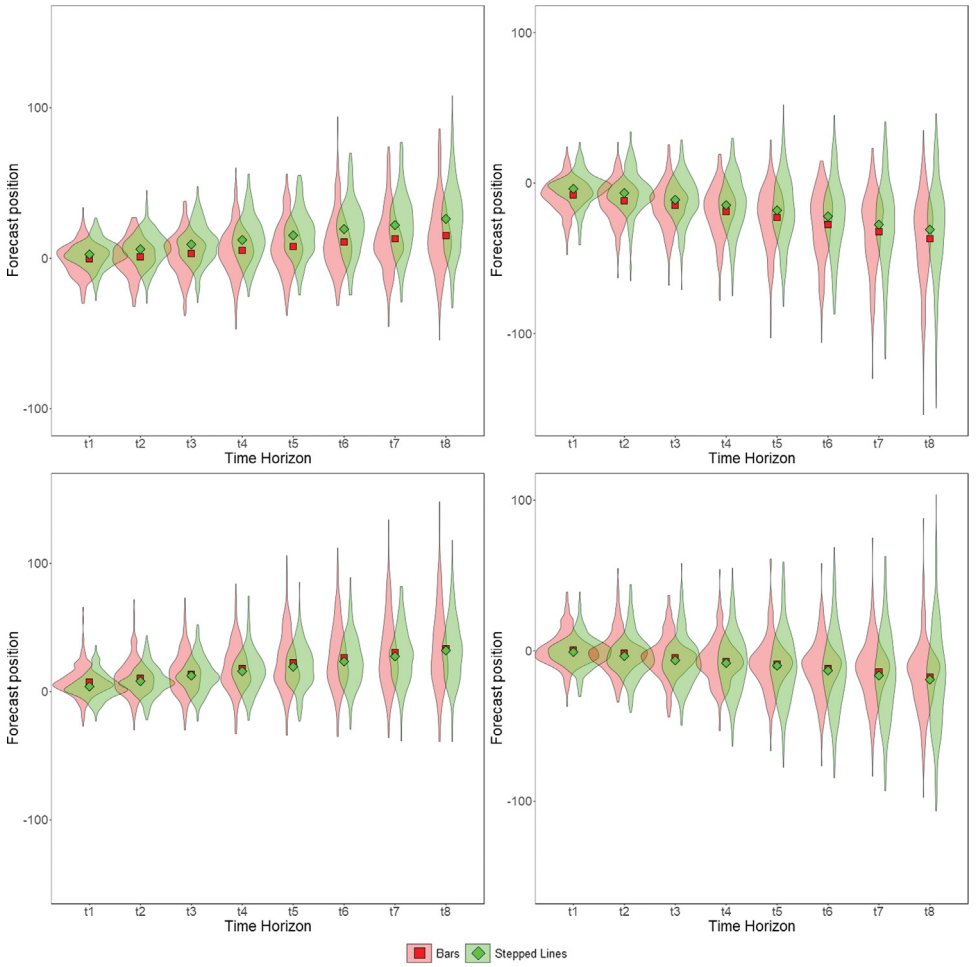

Fig. 6. Mean forecasts for Experiment 3a (Top Left) and 3b (Top Right), comparing bars and stepped lines, and Experiment 4a (Bottom Left) and 4b (Bottom Right), comparing hanging bars and stepped lines.

Table 7
Summary of Effects and interactions in Experiment 3a.

Variable	Degrees of freedom	F ratio	p	η^2_G
Function	**2, 555**	**46.4**	**<.001**	**.105**
Format	**1, 555**	**47.6**	**<.001**	**.057**
Noise	**1, 555**	**30.6**	**<.001**	**.037**
Function x Format	2, 555	0.04	.96	0
Function x Noise	**2, 555**	**18.9**	**<.001**	**.046**
Format x Noise	1, 555	3.80	.052	.005
Function x Format x Noise	**2, 555**	**1.84**	**.16**	**.005**
Time Horizon	**3.4, 1870**	**338**	**<.001**	**.154**
Function x Time Horizon	**6.7, 1870**	**33.9**	**<.001**	**.035**
Format x Time Horizon	**3.4, 1870**	**12.8**	**<.001**	**.007**
Noise x Time Horizon	**3.4, 1870**	**4.95**	**.001**	**.003**
Function x Format x Time Horizon	6.7, 1870	1.47	.18	.002
Function x Noise x Time Horizon	**6.7, 1870**	**6.26**	**<.001**	**.007**
Format x Noise x Time Horizon	3.4, 1870	0.25	.88	<.001
Function x Format x Noise x Time Horizon	6.7, 1870	0.66	.70	.001

Table 8
Summary of Effects and interactions in Experiment 3b.

Variable	Degrees of freedom	F ratio	p	η^2_G
Function	**2, 597**	**61.6**	**<.001**	**.129**
Format	**1, 597**	**8.58**	**.004**	**.010**
Noise	1, 597	2.47	.12	.003
Function x Format	2, 597	1.15	.32	.003
Function x Noise	**2, 597**	**3.79**	**.02**	**.009**
Format x Noise	**1, 597**	**4.77**	**.03**	**.006**
Function x Format x Noise	2, 597	0.53	.59	.001
Time Horizon	**2.3, 1347**	**471**	**<.001**	**.182**
Function x Time Horizon	**4.5, 1347**	**38.6**	**<.001**	**.035**
Format x Time Horizon	2.3, 1347	0.45	.66	<.001
Noise x Time Horizon	**2.3, 1347**	**3.29**	**.03**	**.002**
Function x Format x Time Horizon	4.5, 1347	0.85	.50	.001
Function x Noise x Time Horizon	4.5, 1347	1.97	.09	.002
Format x Noise x Time Horizon	2.3, 1347	1.03	.36	<.001
Function x Format x Noise x Time Horizon	4.5, 1347	1.10	.36	.001

7. Experiment 3b

7.1. Method

7.1.1. Participants

Participants were recruited using the same method as in Experiment 3a from the same participant pool. A total of 892 participants completed the experiment.

7.1.2. Design and procedure

Experiment 3b used the same design and procedure as Experiment 3a, except with downward rather than upward trends.

7.2. Results

Data from 88 participants who had completed one of the other studies reported here were removed, leaving 791 participants, of whom 54% were female, 45% were male, and 1% did not report their gender. The median age was 44 years old (IQR: 24 years). After removing outliers, data from 609 participants entered the analysis (see Fig. 6, top right panel for an overview). The main results were grossly similar to those for Experiment 3a (Table 8).

As in Experiment 3a, there was a main effect of format, but this time no interaction, suggesting here format effects were similar across the eight forecasts. There was also an interaction between format and noise. When examining the effects of format at each level of noise separately, there was no evidence of a format effect with low noise, $F(1, 284) = 0.42$, $p = .52$, $\eta^2_G = .001$, but a clear format effect with high noise, $F(1, 313) = 10.4$, $p = .001$, $\eta^2_G = .023$.

7.3. Discussion

In experiments using upward (Experiment 3a) and downward (Experiment 3b) trends, participants made lower forecasts using bar graphs than using stepped lines, which resembled bar graphs in all respects except for the vertical lines forming the bars. This suggests that differences between lines and bars do not simply arise from differences in presented gradients seen in line graphs, supporting H5.

This leaves two potential accounts. One is that the vertical lines in bar charts draw attention to the area beneath the bar at the expense of attention to the area above the bar. An alternative prediction is that lower forecasts for bars are due to an absolute effect related not to bar height but to physical position on the graph. (One potential speculative account is that bars are seen as inherently more physical than lines and perhaps more subject to physical laws such as 'what goes up must come down.' Or bars have a mass that implicitly drags them down, suggesting that bars should tend to sink towards the bottom of the graph over time.)

In these final experiments, we attempt to distinguish between these two account classes by comparing the same time series as in Experiment 3, but with bars descending from above rather than rising from below (see Fig. 5). We hypothesize:

H6: Participants will shorten bars. In other words, forecasts using bars will, relative to stepped lines, be lower for trends above the *x*-axis and higher for trends that hang below the x-axis

8. Experiment 4a

The design of Experiment 4 was very similar to that of Experiment 3, the major difference being that, in the bar graph conditions, the bars hung down from the top rather than rising from the bottom. For both bars and stepped lines, the *x*-axis was presented at the top of the screen, and the line or bars were presented below. In Experiment 4a, all trends were upward. In Experiment 4b, all trends were downward.

8.1. Method

8.1.1. Participants

Participants were recruited using the same method as in Experiment 1, from the same participant pool. A total of 841 participants completed the experiment.

8.1.2. Design and procedure

As in Experiment 1a, Experiment 4a used a fully between-participants factorial design comprising the following variables: Format (Stepped Lines, Bars), Noise (Low,

Table 9
Summary of Effects and interactions in Experiment 4a.

Variable	Degrees of freedom	F ratio	p	η^2_G
Function	**2, 575**	**56.5**	**<.001**	**.125**
Format	**1, 575**	**4.07**	**.04**	**.005**
Noise	1, 575	1.08	.30	.001
Function x Format	2, 575	3.00	.051	.007
Function x Noise	**2, 575**	**9.35**	**<.001**	**.023**
Format x Noise	**1, 575**	**3.99**	**.046**	**.005**
Function x Format x Noise	2, 575	2.61	.07	.007
Time Horizon	**2.8, 1636**	**564**	**<.001**	**.212**
Function x Time Horizon	**5.7, 1636**	**42.2**	**<.001**	**.039**
Format x Time Horizon	2.8, 1636	0.88	.45	<.001
Noise x Time Horizon	2.8, 1636	1.99	.12	.001
Function x Format x Time Horizon	**5.7, 1636**	**2.25**	**.04**	**.002**
Function x Noise x Time Horizon	**5.7, 1636**	**5.38**	**<.001**	**.005**
Format x Noise x Time Horizon	**2.8, 1636**	**6.09**	**.001**	**.003**
Function x Format x Noise x Time Horizon	**5.7, 1636**	**3.56**	**.002**	**.003**

High), and Function Type (Accelerating, Linear, Decelerating). As before, participants saw a noisy trend of 50 observations and then made a forecast for the next eight observations.

8.2. Results

Data from 74 participants who had completed one of the other studies reported here were removed, leaving 767 participants, of whom 51% were female, 48% were male, and 1% did not report their gender. The median age was 42 years old (IQR: 25 years). After the removal of outliers, 587 participants entered the analysis. The analysis is summarised in Table 9, and overall effects can be seen in Fig. 6, bottom left panel.

There is a complex pattern of interactions involving Function, which will not be discussed in detail here but suggests that the effects of format are potentially contingent on the properties of the Function being forecast. However, there was a three-way interaction between Format, Noise, and Time Horizon, which we explored further. Here, the effect of format was only significant in the Low Noise condition, $F(1, 277) = 11.64$, p <.001 = .031, where there was a significant effect of format at all levels of Time Horizon (Fs = 7.7 – 12.0, ps = .006 – <.001, η^2_G = .027 – .039). There was no significance in the High Noise condition, $F(1, 298) = 0.00$, $p = .99$, in which an effect of format was only seen at the first level of Time Horizon, $F(1, 298) = 7.67$, $p = .006$, $\eta^2_G = .025$, but no other levels (Fs = 0.01 – 1.31, ps = .94 - .25, η^2_G = <.001 – .004). Crucially, where effects of the format were seen, they were in the *opposite* direction to that of Experiment 3, with forecasts from hanging bars higher than those from stepped lines.

9. Experiment 4b

9.1. Method

9.1.1. Participants

Participants were recruited using the same method as in Experiment 4a from the same participant pool. A total of 843 participants completed the experiment.

9.1.2. Design and procedure

Experiment 4b used the same design and procedure as Experiment 4a, except with downward rather than upward trends.

9.2. Results

Data from 86 participants who had completed one of the other studies reported here were removed, leaving 757 participants, of whom 52% were female and 47% were male. The median age was 44 years old (IQR: 24 years). After removing outliers, data from 572 participants remained in the analysis. Results are summarised in Table 10.

Here, unlike Experiment 4a there was no Format x Time Horizon interaction, nor a main effect of Format, suggesting that, more than in earlier studies, the format was not, on average, having a substantial effect on forecasts, as seen in Fig. 6, bottom right panel.

9.3. Discussion

Experiments 4a and 4b help to isolate the locus of the effect by which forecasts are lower from bar graphs than from (stepped or standard) line graphs. Specifically, when participants made forecasts from the same time series as in Experiment 3, but this time with bars hanging down from the top of the graph rather than rising from the bottom, forecasts using bars were no longer lower than those using stepped lines. There was a small but significant reversal for ascending trends in that participants made higher forecasts to hanging bars than stepped lines in the low noise condition. There was a similar qualitative reversal for descending trends, but the difference between forecasts to lines and bars was non-significant. This provides substantial but not unequivocal support for H6.

The findings suggest that the lower forecasts made with bars in earlier experiments are due to participants shortening the height of the bars rather than having the anticipation of lower values in absolute terms. This suggests that accounts based on the asymmetry above and below bars are more compelling than those in which the overall elevation of forecasts is affected by format. There

Table 10
Summary of Effects and interactions in Experiment 4b.

Variable	Degrees of freedom	F ratio	p	η^2_G
Function	**2, 560**	**21.7**	**<.001**	**.055**
Format	1, 560	1.78	.183	.002
Noise	**1, 560**	**35.0**	**<.001**	**.045**
Function x Format	2, 560	2.90	.056	.008
Function x Noise	**2, 560**	**8.70**	**<.001**	**.023**
Format x Noise	1, 560	0.11	.742	<.001
Function x Format x Noise	2, 560	0.41	.667	.001
Time Horizon	**2.4, 1344**	**222**	**<.001**	**.092**
Function x Time Horizon	**4.8, 1344**	**23.6**	**<.001**	**.021**
Format x Time Horizon	2.4, 1344	0.54	.615	<.001
Noise x Time Horizon	**2.4, 1344**	**8.00**	**<.001**	**.004**
Function x Format x Time Horizon	**4.8, 1344**	**5.44**	**<.001**	**.005**
Function x Noise x Time Horizon	**4.8, 1344**	**4.28**	**.001**	**.004**
Format x Noise x Time Horizon	2.4, 1344	1.37	.255	.001
Function x Format x Noise x Time Horizon	4.8, 1344	1.08	.369	.001

are various mechanisms by which this could occur: At a perceptual level, it could be that forecasters'attention is drawn towards the vertical lines included in bar graphs or the implied area within them, and they make forecasts closer to their focus of attention. Alternatively, a more cognitive account might suggest that participants interpret a point inside a bar as more representative or more likely to occur than a similar one outside the bar (see Newman & Scholl, 2012).

The main caveat to these findings is that the complex pattern of three- and four-way interactions among independent variables may indicate that some of these effects, although seen overall, may vary in magnitude and direction across different trend types, noise levels, and so on. Alternatively, it may be that, for Experiments 3 and 4, where we had an unusually high number of extreme forecasts, the process of cell-by-cell outlier removal led to an asymmetry in the removal of spurious responses, which gave some of the more complex effects. If outliers are not removed, the data are much noisier; however, most format effects remain, whereas most higher-level interactions disappear (Supplementary Material, S1).

10. General discussion

Four experiments with over 4000 participants showed that forecasts from trended time series are influenced systematically by the format in which the time series are presented and extrapolated. The clearest finding replicated across Experiments 1, 2, and 3 was that, for the conventional presentation of the same time series data in bar and line charts, forecasts with bars were lower than those with lines (and in Experiment 1, lower than those with points). In most cases, these differences increased as the distance into the future for which the forecast was made increased. This finding was not moderated by the salience of the bars' shading (Experiment 2) or by the fact that bar charts do not contain the gradient information that line charts do (Experiment 3). Experiment 4, although not completely conclusive, indicated that participants' lower forecasts to bars are interpretable as a preference for choosing points closer to the *x*-axis or within the bar rather than for choosing points lower on the *y*-axis in absolute terms. As well as finding the effects of format on general elevation, we also found that participants added more noise to their forecasts with bars than points (Experiment 1).

In addition to these novel findings, we also replicated several previously reported findings. Specifically, we found that participants added noise to their forecasts across all formats and that this noise was higher when the time series itself was noisier (Harvey, 1995). In effect, participants, to some degree, matched the amount of noise in their forecasts to the amount of noise in the time series they experienced. We also saw substantial evidence of trend damping (Fig. 2): For the majority of trends, both upward and downward, participants' forecasts were drawn towards the horizontal (Harvey & Reimers, 2013).

10.1. Format biases with bars

The finding that participants made lower forecasts with bars than with lines and points is perhaps simultaneously surprising and predictable, given recent research. It is surprising since it is non-normative and has substantial implications for unaided judgmental forecasting and, conceivably, situations in which humans judgmentally adjust algorithmic forecasts. In considering real-life forecasting, whether at an individual or organizational level, much of the relevant psychological research has focused on how time series data can lead to suboptimal extrapolation. The exact way time series data have been presented in judgmental forecasting tasks has received relatively little attention (though, for an exception, see Theocharis et al., 2019). It seems clear from these findings that judgmental forecasts can – at least in some circumstances – also be affected by choice of graphical format in which data are presented to forecasters.

This aligns with research on graph comprehension and, more recently, on the within the bar phenomenon (Newman & Scholl, 2012; Okan, Garcia-Retamero, Cokely, & Maldonado, 2018) which shows that format affects how graphical data are cognitively represented. It is therefore not at all surprising that format also affects the way that people extrapolate from time series.

10.2. Adding noise to forecasts

Across all experiments, we found that, although normative forecasts to the time series would be linear or near-linear extrapolations of the time series at the gradient around the final observation, participants added noise to their forecasts and that this noise was larger when the time series were noisier. This is not surprising: When evaluating stochastically generated outcomes, people rate the probability of occurrence as a function of the perceived randomness or unpredictability of the trend (Kahneman & Tversky, 1972; Reimers, Donkin, & Le Pelley, 2018). This suggests the use of a representativeness heuristic: Participants appear to produce forecasts that have properties that are representative of the time series they are evaluating, in particular noisiness, even though for forecasting, this leads to less accurate performance. One way of countering this propensity, beyond using scatterplots rather than bar charts, might be to alter the instructions to encourage participants to aggregate noise away: "Imagine this plays out a hundred times. Each time it will be slightly different. You need to indicate where you think the middle of all those hundred outcomes would be for each of the time points".

The explanation for participants' adding more noise when data were displayed as bars rather than points must be more speculative, but two potential accounts stand out. The first is the salience of the noise, which to us seems stronger in the bars condition than the points condition (see Fig. 1). The second is pragmatics: Scatterplots are frequently presented to allow people to add their own mental or physical best-fit lines to make predictions. Indeed, in the UK education system, children take maths exams in which they have to linearly interpolate or extrapolate data presented as scatterplots. As such, the presentation of data as points may have cued participants to look more at the underlying trend.

10.3. Implications

There are significant potential implications for forecasting practitioners and financial regulators. If format changes can systematically manipulate an individual's forecasting of a time series, then the graph designer carries some responsibility for the graph's interpretation.

One positive implication of this is that there is potential for using format effects to overcome other biases seen in forecasting. In situations where biases lead to judgmental forecasts being too high, for example, as the result of an optimism bias or the damping of a downward trend, using bars rather than lines may help correct these biases.

Conversely, there are more negative implications: people who want others to interpret a graph in a particular way can improve the chances of that kind of interpretation by their choice of format. For example, people's expectations about the future performance of a fund might be somewhat more positive if the fund manager presents the historical performance as a line graph rather than a bar graph. A government report might make people happier with crime statistics trends if presented as bar graphs rather than line graphs.

Although the main effects of the format were robust across multiple levels of noise and function type, the higher-level interactions found in some of the studies suggest that the precise magnitude (and sometimes direction) of different format effects may differ across different contexts. Further research on the factors affecting format effects – including the precise way each graph format is presented – would be merited before any firm conclusions are drawn. However, these findings demonstrate that format has a clear effect both on the absolute values of forecasts and on the noise added to forecasts, and so should be considered in the design of judgmental forecasting processes and systems.

10.4. Limitations

Although the studies reported here had some unique strengths – a large number of participants, a degree of incentivization, and fully between-participants designs – some limitations should be taken into account when considering the generalizability of the results. First, the time series were artificial and trended. This allowed us to control the extraneous factors that could have affected forecasting. Still, with more ecological time series, other factors could overshadow or moderate the effects observed here. From the higher-order interactions found in some of the studies, the magnitude of format, shading, and noise effects on forecasts may vary by the underlying function being forecast. Similar variability may be seen with more ecological trends and forecasts concerning other outcome types; clearly, identifying more systematically the factors that modulate format and other effects would be worthy of further research.

Relatedly, the effect sizes for effects involving format were in the small-to-medium range, so there is a question of the real-world relevance of these findings. Effect sizes are driven not only by the differences in means between distributions but also by differences in the variances of the distributions. We suspect that our one-shot online methodology may have led to larger response variability (thereby reducing effect sizes) than a more traditional study in which participants complete many forecasting trials. It would be fruitful to compare the effects of format with other factors that might influence judgmental forecasting, including the underlying trend. However, we note that in the studies reported here, the effect of format was sometimes larger than that of the underlying trend, even though the trends differed relatively substantially.

Second, although we used several different presentation formats, each format was implemented in a single way. It is uncertain how these effects would generalize to other methods of presenting bars, lines, and points. It is perhaps most noticeable for the Bars condition, in which, to maximize presentational consistency with the line and point conditions, bars were tightly packed with no gaps between them. Although this kind of bar chart does occur in real situations, it is more common to see wider bars with gaps between them. Whether this modulates the 'lower forecasts to bars' effect is an open empirical question.

There may also be an effect of expertise. These format effects were found in naïve forecasters in a domain for which they had no information or experience beyond the time series. More experienced forecasters might show a different pattern of format effects (cf. Cardinaels, 2008). Similarly, format-specific expertise is likely to have an impact – people used to interpreting bar charts are likely to show different format effects from those used to interpreting scatterplots.

In our task, participants made their forecasts by clicking a point on the graph displayed on their screen. Would we have obtained different results if we had asked them to type their forecasts into a box below the screen? Responses are generally more accurate (and faster) when stimulus–response compatibility is high (Proctor & Reeve, 1989). To ensure high stimulus–response compatibility in our studies, we – like most previous researchers on judgmental forecasting – required forecasts to be added to the graph used to present the data series. If we had presented the data series as a table of numbers, high stimulus–response compatibility would have been achieved by asking forecasters to type in numbers to add to the figures in the table. If we had asked people to type in numbers to make forecasts from graphically presented data, stimulus–response compatibility would have been low. From previous research (Proctor & Reeve, 1989), we would have expected this to reduce response accuracy by adding noise to forecasts without affecting the observed biases. In other words, format effects would have persisted but would have been overlaid by more data noise.

It is possible to argue that line and point graphs are more compatible with continuous variables, whereas bar graphs are more compatible with discrete variables. This would lead us to expect that forecasting biases are *less* with bar graphs than with line or point graphs when plots represent discrete rather than continuous variables. However, in our experiments, people forecast the 'Number of Sales,' a discrete variable. As we found that biases were *greater* with bar graphs, our findings are inconsistent with data type/format type compatibility effects.

10.5. Future research

In our task, forecasters made a set of eight forecasts from a single series for periods 1–8 ahead. Would format effects have been different if a series of one-period ahead forecasts had been made in which the series was updated after each of those forecasts, thereby providing immediate outcome feedback? Two points should be made here. First, in most domains (e.g., demand forecasting), practitioners do not receive feedback. This is because forecasts are either not retained or because, if they are, they are not compared with outcomes. Second, as Niu and Harvey's (2022) review demonstrated, no previous work, until the studies they report, had compared the effects of providing simple outcome feedback with the effects of not providing that feedback. Their studies showed no effect of providing outcome feedback on overall error measures. However, this was because, though outcome feedback almost halved constant error (bias), it increased variable error (noise) by a corresponding amount. Thus, we expect that outcome feedback would reduce but not eliminate the format effects that we have reported because of its effect on the bias. However, further research is needed to confirm this.

Our experiments' data points were independently distributed around a mean or trend line. Would the effects on format have been different if these points had been serially dependent? Reimers and Harvey (2011) showed that the positioning of people's forecasts is consistent with their overestimating the first order autocorrelation in data series when it is low (e.g., 0.00) but underestimating the autocorrelation when it is high (e.g., 0.80). More recently, Theocharis et al. (2019) demonstrated that this effect of serial dependence was present in both line and point graphs, but it was greater with line graphs. They argued that the connecting lines between points in line graphs emphasized the potential interdependence between successive points. If we had manipulated serial dependence in our experiments, we would have expected to replicate their finding that its effects are greater in line than in point graphs. Theocharis et al. (2019) did not include bar graphs in their experiments. However, if their account is correct, we expect the effects of series autocorrelation to be less in bar graphs than in line graphs because there are no explicit links between successive bars. Further research is needed to confirm this.

10.6. Conclusion

In summary, the extent to which format effects are likely to impact people's real choices will depend not only on the factors we have examined but also on several as-yet-unexplored variables. This includes different ways of presenting line and bar charts, different factors relating to presentation and response mode, and possible effects of more contextualized realistic types of time series. Nevertheless, the reported findings provide early evidence that formats are not interchangeable in all situations and that attention to the choice of format for conveying time series information is important.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2022.11.003.

References

Andreassen, P. B., & Kraus, S. J. (1990). Judgmental extrapolation and the salience of change. *Journal of Forecasting*, *9*, 347–372.

Bolger, F., & Harvey, N. (1993). Context-sensitive heuristics in statistical reasoning. *The Quarterly Journal of Experimental Psychology Section A*, *46*, 779–811.

Bolger, F., & Harvey, N. (1998). Heuristics and biases in judgmental forecasting. In G. Wright, & P. Goodwin (Eds.), *Forecasting with judgment* (pp. 113–137). New York USA: John Wiley & Son.

Cardinaels, E. (2008). The interplay between cost accounting knowledge and presentation formats in cost-based decision making. *Accounting, Organizations, and Society*, *33*(6), 582–602.

Carpenter, P. A., & Shah, P. (1998). A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied*, *4*(2), 75–100.

Chapman, G. B., & Johnson, E. J. (1999). Anchoring, activation, and the construction of values. *Organizational Behavior and Human Decision Processes*, *79*(2), 115–153.

Correll, M., & Heer, J. (2017). Regression by eye: Estimating trends in bivariate visualizations. In *Proceedings of the 2017 CHI conference on human factors in computing systems* (pp. 1387–1396).

Desanctis, G., & Jarvenpaa, S. L. (1989). Graphical presentation of accounting data for financial forecasting: An experimental investigation. *Accounting, Organizations and Society*, *14*, 509–525.

Fildes, R., & Goodwin, P. (2007). Good and bad judgment in forecasting: Lessons from four companies. *Foresight*, (Fall 2007), 5–10.

Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting*, *25*(1), 3–23.

Fildes, R., & Petropoulos, F. (2015). Improving forecast quality in practice. *Foresight*, (Winter 2015), 5–12.

Goodwin, P., & Wright, G. (1993). Improving judgmental time series forecasting: A review of the guidance provided by research. *International Journal of Forecasting*, *9*, 147–161.

Goodwin, P., & Wright, G. (1994). Heuristics, biases and improvement strategies in judgmental time series forecasting. *Omega*, *22*, 553–568.

Harvey, N. (1995). Why are judgments less consistent in less predictable task situations? *Organizational Behavior and Human Decision Processes*, *63*, 247–263.

Harvey, N. (2007). Use of heuristics: Insights from forecasting research. *Thinking and Reasoning*, *13*, 5–24.

Harvey, N., & Bolger, F. (1996). Graphs versus tables: Effects of data presentation format on judgemental forecasting. *International Journal of Forecasting*, *12*, 119–137.

Harvey, N., & Reimers, S. (2013). Trend damping: Under-adjustment, experimental artifact, or adaptation to features of the natural environment? *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *39*, 589–607.

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

Lawrence, M., Goodwin, P., O'Connor, M., & Önkal, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting*, *22*, 493–518.

Lawrence, M., & Makridakis, S. (1989). Factors affecting judgmental forecasts and confidence intervals. *Organizational Behavior and Human Decision Processes*, *43*, 172–187.

Meyer, J., Shamo, M. K., & Gopher, D. (1999). Information structure and the relative efficacy of tables and graphs. *Human Factors*, *41*, 570–587.

Newman, G. E., & Scholl, B. J. (2012). Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar bias. *Psychonomic Bulletin & Review*, *19*(4), 601–607.

Niu, X., & Harvey, N. (2022). Outcome feedback reduces overforecasting of inflation and overconfidence in forecasts. *Judgment and Decision Making*, *17*(1), 124–163.

O'Connor, M., Remus, W., & Griggs, K. (1997). Going up–going down: How good are people at forecasting trends and changes in trends? *Journal of Forecasting*, *16*, 165–176.

Okan, Y., Garcia-Retamero, R., Cokely, E. T., & Maldonado, A. (2018). Biasing and debiasing health decisions with bar graphs: Costs and benefits of graph literacy. *Quarterly Journal of Experimental Psychology*, *71*(12), 2506–2519.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods*, *8*(4), 434–447.

Petropoulos, F., Kourentzes, N., Nikolopoulos, K., & Siemsen, E. (2018). Judgmental selection of forecasting models. *Journal of Operations Management*, *60*(1), 34–46.

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

Proctor, R. W., & Reeve, T. G. (1989). Stimulus–response compatibility: An integrated perspective. In *Elsevier advances in psychology (vol. 65)*.

Reimers, S., Donkin, C., & Le Pelley, M. E. (2018). Perceptions of randomness in binary sequences: Normative, heuristic, or both? *Cognition*, *172*, 11–25.

Reimers, S., & Harvey, N. (2011). Sensitivity to autocorrelation in judgmental time series forecasting. *International Journal of Forecasting*, *27*, 1196–1214.

Reimers, S., & Stewart, N. (2007). Adobe Flash as a medium for online experimentation: A test of reaction time measurement capabilities. *Behavior Research Methods*, *39*(3), 365–370.

Reimers, S., & Stewart, N. (2015). Presentation and response timing accuracy in Adobe Flash and HTML5/JavaScript Web experiments. *Behavior Research Methods*, *47*(2), 309–327.

Schonlau, M., & Peters, E. (2012). Comprehension of graphs and tables depend on the task: empirical evidence from two web-based studies. *Statistics, Politics, and Policy*, *3*(2), Article 5.

Shah, P., & Hoeffner, J. (2002). Review of graph comprehension research: Implications for instruction. *Educational Psychology Review*, *14*, 47–69.

Simcox, W. A. (1984). A method for pragmatic communication in graphic displays. *Human Factors*, *26*(4), 483–487.

Speier, C. (2006). The influence of information presentation formats on complex task decision-making performance. *International Journal of Human-Computer Studies*, *64*, 1115–1131.

Theocharis, Z., Smith, L. A., & Harvey, N. (2019). The influence of graphical format on judgmental forecasting accuracy: Lines versus points. *Futures & Foresight Science*, *1*, e7.

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

Tversky, B., Zacks, J., Lee, P. U., & Heiser, J. (2000). Lines, blobs, crosses, and arrows: Diagrammatic communication with schematic figures. In M. Anderson, P. Cheng, & V. Haarslev (Eds.), *Theory and application of diagrams* (pp. 221–230). Berlin, Germany: Springer.

Vessey, I., & Galletta, D. (1991). Cognitive fit: An empirical study of information acquisition. *Information Systems Research*, *2*, 63–84.

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt II. In *Psycologische forschung (vol. 4)* (pp. 301–350).

Zacks, J., & Tversky, B. (1999). Bars and lines: A study of graphic communication. *Memory and Cognition*, *27*, 1073–1079.