Cognitive Psychology 128 (2021) 101406

Contents lists available at ScienceDirect

Cognitive Psychology

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

Can humans perform mental regression on a graph? Accuracy and bias in the perception of scatterplots

Lorenzo Ciccione [a,b,c,*], Stanislas Dehaene [b,c]

[a] University Paris Sciences Lettres (PSL), 60 rue Mazarine, 75006 Paris, France
[b] Cognitive Neuroimaging Unit, CEA, INSERM, Université Paris-Saclay, NeuroSpin Center, 91191 Gif/Yvette, France
[c] Collège de France, Université Paris Sciences Lettres (PSL), 11 Place Marcelin Berthelot, 75005 Paris, France

ARTICLE INFO

Keywords:
Graph perception
Graphicacy
Trend judgment
Regression
Extrapolation
Scatterplot
Cognitive bias

ABSTRACT

Despite the widespread use of graphs, little is known about how fast and how accurately we can extract information from them. Through a series of four behavioral experiments, we characterized human performance in "mental regression", i.e. the perception of statistical trends from scatterplots. When presented with a noisy scatterplot, even as briefly as 100 ms, human adults could accurately judge if it was increasing or decreasing, fit a regression line, and extrapolate outside the original data range, for both linear and non-linear functions. Performance was highly consistent across those three tasks of trend judgment, line fitting and extrapolation. Participants' linear trend judgments took into account the slope, the noise, and the number of data points, and were tightly correlated with the *t*-test classically used to evaluate the significance of a linear regression. However, they overestimated the absolute value of the regression slope. This bias was inconsistent with ordinary least squares (OLS) regression, which minimizes the sum of square deviations, but consistent with the use of Deming regression, which treats the x and y axes symmetrically and minimizes the Euclidean distance to the fitting line. We speculate that this fast but biased perception of scatterplots may be based on a "neuronal recycling" of the human visual capacity to identify the medial axis of a shape.

1. Introduction

What is a graph? According to the Cambridge Dictionary of English, a graph is *a picture that shows how two sets of information or variables are related.* Edmund Halley, the famous English astronomer, is generally considered the inventor of the line graph in 1686, although the first graphical representation of real-world data is attributed to the Scottish engineer William Playfair in 1786 (for an historical summary, see Spence, 2006). Although graphical representations of scientific data are a recent cultural invention, they have become ubiquitous in our life. In the past two centuries, graphs have become the elective tool of various professionals, such as journalists, statisticians, doctors and scientists, who commonly use them to represent data that are too numerous or complex to be represented through plain text or explicit mathematical notation. The reason for this success is that effective graphs can convey, in a quick and efficient manner, a large set of complex data and their relationships (Tufte, 2001). Nevertheless, the scientific investigation of graph perception is far from exhaustive, especially if compared to the study of the two most famous human cultural inventions, reading and mathematics. Here, through several psychophysical experiments, we aim to bridge this gap and begin to characterize, in

* Corresponding author at: Cognitive Neuroimaging Unit, CEA, INSERM, Université Paris-Saclay, NeuroSpin Center, 91191 Gif/Yvette, France
E-mail address: lorenzo.ciccione@cri-paris.org (L. Ciccione).

https://doi.org/10.1016/j.cogpsych.2021.101406
Received 6 January 2021; Received in revised form 17 June 2021; Accepted 18 June 2021

some detail, the perception of scatterplots in educated human adults.

Graphs, written words, numbers and equations share several features. First, they are symbolic representations based on shared conventions. Second, they take advantage of the speed and capacity of the human visual channel to allow for the fast transmission of complex information. Third, they are cultural inventions that require considerable learning experience and are commonly taught at school. To further underline the parallelism with literacy and numeracy, the ability to understand graphs has been named *graphicacy* (Balchin & Coleman, 1966).

According to the scientific literature on this topic, mastering graphicacy requires three steps, which have been metaphorically defined as (1) reading *the* data, (2) reading *between* the data and (3) reading *beyond* the data (Curcio, 1987; Friel, Curcio, & Bright, 2001). The first step consists in extracting data by visually inspecting the elements drawn in the graph; the second step requires inferring the mathematical relation between the variables; the third step refers to the ability to make a numerical forecast based on the graph's underlying trend. This three-steps classification is particularly useful to organize the heterogeneous corpus of studies about graph perception and comprehension.

1.1. Data extraction

The first step, data extraction or "reading the data", requires being able to recognize simple perceptual aspects of the graph and to understand the symbols that it comprises, in order to extract relevant numerical data. This process has also been termed *visual decoding* (Cleveland & McGill, 1984). Through a series of behavioral studies, Cleveland and McGill classified the different visual tools used in a graph in terms of the complexity of data extraction. For instance, they found that assessing the position of points along an axis is usually easier than determining the slope of a line (Cleveland & McGill, 1985) and that comparing the slopes of two lines is easier when their mid angle is close to 45° (Cleveland & McGill, 1987). Ways of improving graph readability and visual decoding have also been proposed. Edward Tufte, from an intuitive graphic-design perspective, suggested that the ratio of data to ink should be maximized (Tufte, 2001). Poulton (1985) showed that participants' accuracy at detecting the position of data points in a scatterplot can be improved if the x and y axes are presented twice, on both sides of the graph (note that this recommendation conflicts with Tufte's axiom).

Color and brightness can also affect the ease of data extraction from a graph (Vanderplas, Cook, & Hofmann, 2020). Indeed, most Gestalt principles of perceptual organization play a crucial role in data extraction (Kosslyn & Kosslyn, 2006): for example, data points that are closer in space are more likely to be perceived as grouped (Ciccione & Dehaene, 2020), and vertical and horizontal lines are easier to discriminate than oblique ones (Appelle, 1972). In this respect, graph perception probably relies on universal mechanisms of point, line and angle perception. An interesting and largely open question is to what extent training in graphicacy modifies these perceptual mechanisms. As we will argue in the discussion, graphicacy probably represents a novel instantiation of "neuronal recycling" (Dehaene, 2005; Dehaene & Cohen, 2007), according to which cultural inventions such as written words and numbers efficiently exploit and repurpose evolutionary older cognitive systems to a novel use.

1.2. Inference of the mathematical function

The second step, inference or "reading between the data", refers to the ability to infer the nature of the mathematical function that relates the data on the graph. Most studies in this field have focused on a specific paradigm of "function learning", the ability to learn the functional mapping between a set of input values and a set of output values. However, these studies did not primarily focus on graphs, but mostly on the sequential presentation of pairs of input/output data. In such paradigms, there is a learning phase during which participants slowly infer (with or without corrective feedback) the nature of the relation between paired stimuli, the data being presented for instance as horizontal bars (one for the input value, and another for the output value of the function given that input). When participants were asked to generalize from new input stimuli, they often gave evidence of having correctly learned and applied the function that linked the input and output values (Bott & Heit, 2004; Carroll, 1963; DeLosh, McDaniel, & Busemeyer, 1997). Human function learning has been modelled as a combination of two processes: explicit function estimation and associative learning based on similarity (Lucas, Griffiths, Williams, & Kalish, 2015). This model successfully accounts for various empirical observations, such as the difficulty to learn periodic functions (Kalish, 2013), the human bias towards positive linear trends (Kalish, Griffiths, & Lewandowsky, 2007; Mcdaniel & Busemeyer, 2005), and the knowledge partitioning effect in function learning (Kalish, Lewandowsky, & Kruschke, 2004; Lewandowsky, Kalish, & Ngang, 2002), according to which human adults can learn and hold multiple parcels of knowledge about functional relationships, depending on stimulus' context.

While a growing body of literature focused on the cognitive mechanisms of function learning, few studies investigated the capacity to infer functions from classic graphical stimuli (such as line plots or scatterplots drawn on a Cartesian plane). A notable exception is the work of Schulz, Tenenbaum, Duvenaud, Speekenbrink, and Gershman (2017), who found that human adults could successfully interpolate and extrapolate sophisticated functions (such as a sinusoidal function with an increasing amplitude) from a plot of their graph. Schulz et al. suggested that this ability reflected the existence of a compositional grammar of functions, which allows human adults to understand complex functions as the composition of a small repertoire of simpler ones. Other authors have shown that, when exposed to a noisy scatterplot, participants tend to interpolate functions with a lower polynomial degree than the real one (Little & Shiffrin, 2009) and their subjective ability to interpolate is negatively affected by increasing levels of noise (Schulz, Tenenbaum, Reshef, Speekenbrink, & Gershman, 2015). Other studies showed that participants could, in a slow and reflexive manner, fit a linear function to a given scatterplot after receiving formal training on statistical regressions (Mosteller, Siegel, Trapido, & Youtz, 1981). Human adults may even adjust quadratic and trigonometric functions to an underlying scatterplot, once they are precisely informed

about the nature of each curve (Correll & Heer, 2017).

Closer to the present research, the perceived correlation in a scatterplot has been investigated as well. When asked to judge the degree of association between two variables in a scatterplot, participants tend to underestimate it (Strahan & Hansen, 1978) and this underestimation is higher for regression slopes further from a 45° orientation (Bobko & Karren, 1979). Participants' ability to compare scatterplots with different correlation coefficients follows Weber's law (Harrison, Yang, Franconeri, & Chang, 2014; Rensink & Baldridge, 2010) and they perceive datasets as more highly correlated if the axes are scaled in order to make the underlying function's slope appear steeper (Beattie & Jones, 2002; Cleveland, Diaconis, & Mcgill, 1982). Participants' performance at extracting the regression's slope is also affected by the localization of the scatterplot: positive slopes are more easily detected if the scatterplots are presented on the right of the visualization area, and the opposite is true for negative slopes (Parrott et al., 2014), a bias consistent with the Spatial-Numerical Association of Response Codes (SNARC) effect (Dehaene, Bossini, & Giraux, 1993).

1.3. Numerical forecasting

The third step, numerical forecasting or "reading beyond the data", is the capacity to make numerical predictions and forecasts based on the data presented, i.e. to extrapolate beyond the existing data range. Given its many concrete applications, it is not surprising that the majority of studies in this area have been conducted by researchers in finance, economics and politics. Studies on this topic are too diverse in methods and purposes to be reviewed here. Most pertinent to the present work is a general tendency to underestimate future data points if they are based upon a non-linear, positively accelerated function (Lawrence, Goodwin, O'Connor, & Önkal, 2006; W. Wagenaar & Sagaria, 1975). The authors speculated that, in a real-world context, most trends do not keep growing at the same steady rate and, knowing this, participants might be conservative in their predictions. Better extrapolation performance has been obtained through experimental designs involving numerical values instead of scatterplots (Lawrence & Makridakis, 1989). The noise of the dataset was also found to affect extrapolation performance (Harvey, Ewart, & West, 1997), and participants seemed to add noise to their extrapolations, as if they were attempting to make their predictions more consistent with a noisy forecast (Bolger & Harvey, 1993).

1.4. Current goal: Studying the psychophysics of graph perception

Overall, this heterogeneous set of studies, which vary considerably in both methods and research questions, do not provide a thorough psychophysical investigation of human abilities to detect trends in noisy graphical representations, in the absence of any reference to the graph's meaning, context, or underlying function. The vast majority of these studies used very few stimuli, did not systematically explore the multiple parameters of the graphs, provided considerable training and/or background information about the underlying functions, left considerable time for subjects to inspect the data and strategize about the task, and none of them measured the time needed to extract basic information from a graph, for instance to perform a simple trend judgment.

Here, our goal was to begin to fill these gaps. We ran a series of four behavioral experiments with full factorial designs, with the purpose of studying human accuracy, response times and biases in the visual extraction and inference of statistical information from a graph. For simplicity, we focused on one of the simplest graphical representations, the scatterplot. This choice was made to minimize the effects of learning and memory, which, as pointed out by several authors (Little & Shiffrin, 2009; Lucas et al., 2015) were certainly at play in the aforementioned classic studies of function learning, mostly based on long training sessions with serial presentations of input/output pairs. Indeed, one of the most remarkable properties of the scatterplot is its ability to simultaneously represent, in a single graph, a very large data set. By flashing a scatterplot on each trial and asking participants to make a simple decision about it, we brought classical methods of psychophysics and mental chronometry to the study of human graph perception and inference of mathematical functions. Specifically, our research aimed to answer four empirical questions:

1) Can human adults perform a fast judgment of the *trend* underlying a noisy scatterplot, i.e. understand whether the data is increasing or decreasing? Which factors affect participants' accuracy and response times in such a task? Do participants perform a mental computation akin to the Pearson coefficient of correlation? Do they behave in a near-optimal manner, correctly taking into account the number of data points and their noise level, and ultimately basing their decisions on a computation similar to the *t*-test that a statistician would use to detect if a significant positive or negative trend is present?
2) Can human adults *fit a line* to a noisy scatterplot? Which algorithm do they use to respond? Is it similar to a linear regression? Is their slope estimate close enough to the one that a statistician would compute? Is it biased?
3) Can human adults *extrapolate* a new data point from the noisy scatterplot of a linear function? If so, do they use the same statistical procedure as in the slope-estimation task, and are they affected by the same biases?
4) Can human adults perform a *non-linear extrapolation* of scatterplots for complex functions such as piecewise linear, quadratics, or sinusoids? Can they do so even without previous training, nor any prior knowledge of which functions will be used?

2. Experiment 1: Trend judgment

In this first experiment, we tested if human adults can perform a fast, intuitive judgment of whether a scatterplot of data shows an increasing or decreasing trend. We generated the graphs according to the hypotheses of classical linear regression ("ordinary least squares"): the values on the ordinate (called y_i) were a linear function of the values on the abscissa (called x_i) plus independent Gaussian noise ($y_i = \alpha x_i + \varepsilon_i$, where ε_i are random numbers independently drawn from a normal distribution centered on zero and

with standard deviation σ). We varied orthogonally three parameters of the graphs: the slope of the linear trend (α); the number of points (*n*); and the standard deviation of the noise (σ).

This experimental design was chosen because it allowed to compare the performance of human participants with a normative model of decision making in this task. As further detailed in the appendix, classical statistical theory predicts that the optimal decision should be determined by a simple *t* test, similar to the one that statisticians use to test for the presence of a positive or negative linear trend. The theory further predicts that responses should be a sigmoidal function of the *t* value, and that the response time should be a decreasing, convex upward function of the absolute deviation of the *t* value from zero. The sole dependence of decisions on the *t* value also implied that decision difficulty should vary significantly with all three of the manipulated graph parameters (n, σ and α), because all of them influence the statistical *t* test: the *t* test varies positively and linearly with the slope α, positively with the number of points (as the square root of $n-2$), and inversely with the noise level σ. Finally, the theory predicts that the effects of these variables should be jointly subsumed by an effect of the *t* value on behavior.

2.1. Methods

2.1.1. Participants

10 participants were recruited (age: 23.9 ± 1.5, 4 females, 6 males). All participants had normal or corrected to normal vision, no medical history of epilepsy, were right-handed, and did not take psychoactive drugs. They all signed an informed consent and were paid 5 euros for their participation. The experimental session lasted approximately 30 min. The experimental procedure was approved by the local ethical committee.

2.1.2. Experimental design

Each participant was presented with 672 scatterplots and, for each of them, was asked to decide, as fast as possible, if the dataset was increasing or decreasing. Each scatterplot was the graphical representation of a dataset that was generated randomly, independently for each participant, using a linear equation plus noise (see below). The design was a full factorial design where we varied the number of points ($n = 6$, 18, 38 or 66), the standard deviation of the noise ($\sigma = 0.05$, 0.1, 0.15 or 0.2), and the slope of the underlying linear trend ($\alpha = -0.1875, -0.125, -0.0625, 0, +0.0625, +0.125$ or $+0.1875$), for a total of $4 \times 4 \times 7 = 112$ combinations. The values of *n* were chosen so that $\sqrt{n-2}$, which is the value that enters in the *t*-test for the significance of a regression, was linearly distributed ($\sqrt{n-2} = 2$, 4, 6 or 8). The other factors were selected after piloting in order to avoid excessive difficulty as well as ceiling effects; specifically, we chose relatively high levels of noise and relatively small levels of α in order to make the task non-trivial. The 112

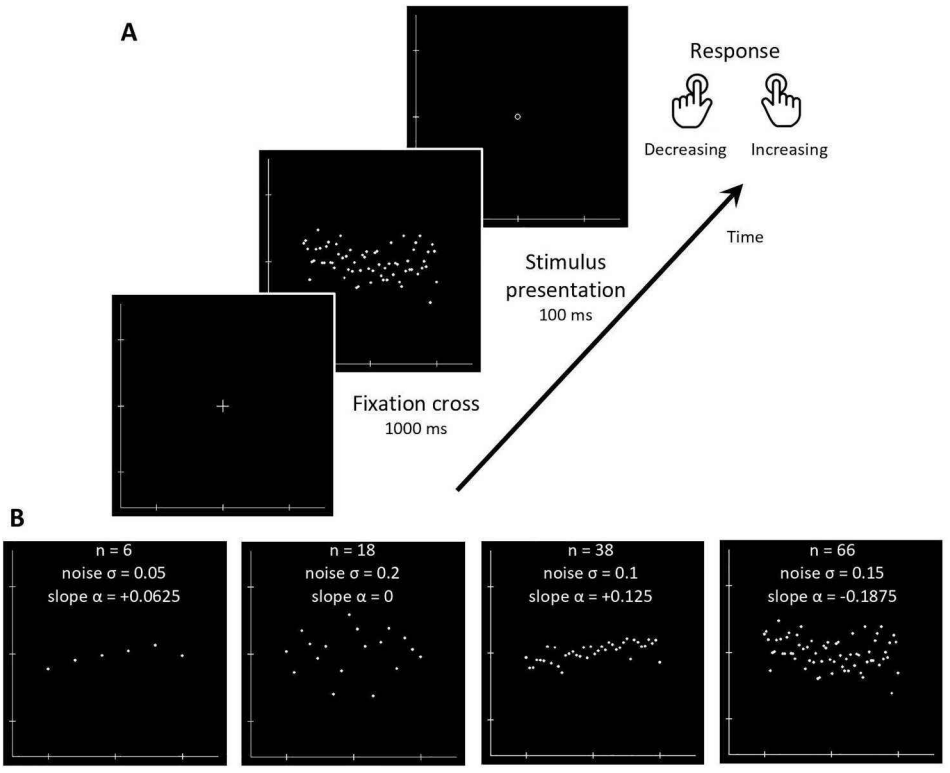

Fig. 1. Experimental design and stimuli for experiment 1. A, task: subjects were presented with a simple scatterplot, generated by a linear function plus noise, with a variable number of data points, and were asked to judge if the trend was ascending or descending. B, examples of stimuli.

combinations of parameter values were randomly presented to each participant in each of the 6 experimental blocks, for a total of 672 trials per participant.

2.1.3. Procedure

The participants were invited to sit on a fixed chair with their head at a distance of 50 cm from the screen. As illustrated in Fig. 1A, a fixation cross first appeared for 1000 ms, immediately followed by the flashing of a scatterplot for 100 ms, and then a fixation circle of 1 cm diameter at the center of the screen; the participants were informed that the circle marked the onset of the response window. Participants were asked to respond as fast and as accurately as possible. Half of them responded by pressing with their right index on a key (signaled with a ⇑ sticker) on the right side of the keyboard if they thought that the trend in the scatterplot was increasing; and, conversely, they pressed with their left index on a key (signaled with a ⇓ sticker) on the left side of the keyboard if they thought that the trend in the scatterplot was decreasing. The opposite response configuration was presented to the other half of the participants. Once they gave their answer, a fixation cross appeared again for 1000 ms, inviting the participants to concentrate on the center of the screen before a new stimulus appeared. As mentioned, the task was divided into 6 blocks of 112 trials; the duration of each block was ~4 min. After each block, the participants could take a short break and received feedback on the total number of correct responses they gave in that block. Before the beginning of the actual experiment, 25 practice trials were run under the supervision of the researcher, in order to control for the correct execution of the task.

2.1.4. Stimuli.

Each scatterplot comprised two unlabeled lines denoting the x and y axes (which remained on screen for the duration of the experiment), each marked with three small ticks at the values 0, 0.5, and 1 (see Fig. 1; those numbers were arbitrary and were not shown to the participants). The dots' coordinates were calculated by a Python program as follows. First, the algorithm used the desired number of dots, *n*, to generate the x values (denoted x_i) such that they ranged from 0 to 1 and were equally spaced. Thus, for example, for all configurations having 6 points, the x values were always [0, 0.2, 0.4, 0.6, 0.8, 1]. The y coordinates were then determined according to the following equation: $y_i = \alpha\, x_i + \varepsilon_i$, where α is the prescribed slope and the ε_i are random numbers independently drawn from a normal distribution centered on zero and with standard deviation σ. If, occasionally, a point took a particularly high or small y coordinate ($y < -0.27$ or $y > 1.27$), which would have exceeded the boundaries of the y axis, the algorithm was reinitialized for that

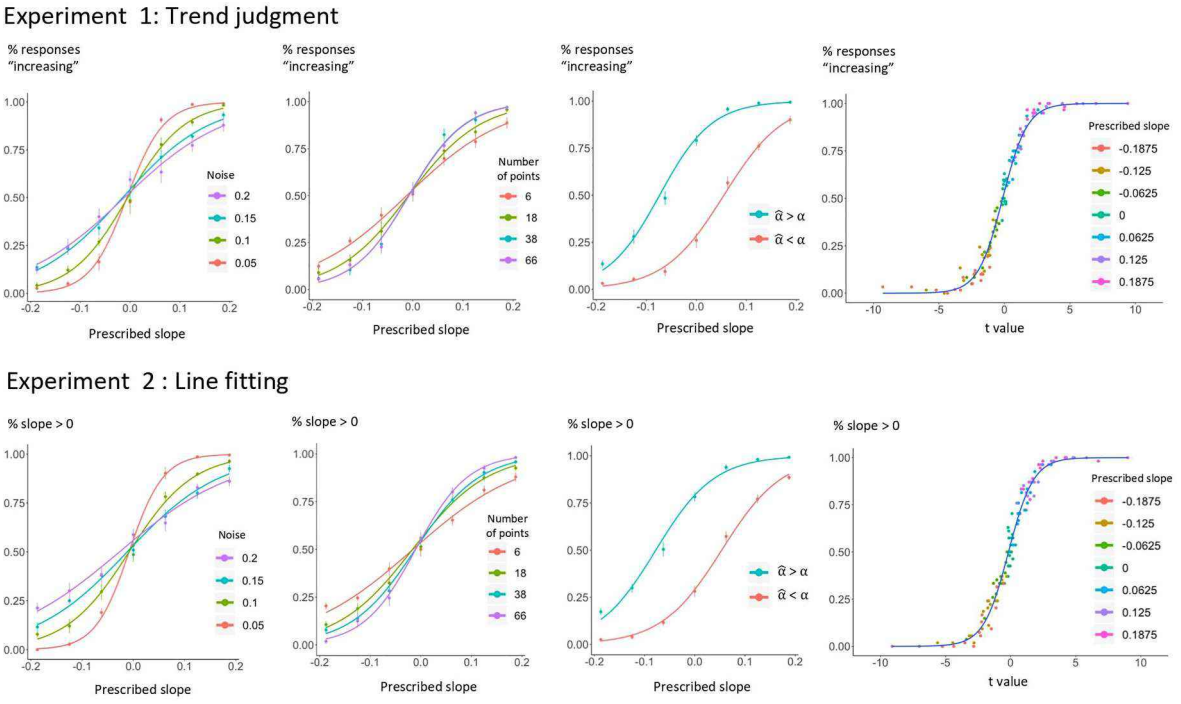

Fig. 2. Accuracy of human subjects in judging whether a noisy scatterplot is increasing or decreasing. In experiment 1, participants made this judgment directly by giving a binary response, and in experiment 2 they fitted a line to the scatterplot, and we categorized its slope as > or <0. In both cases, the percentage of "increasing" responses is affected by the prescribed slope of the graph (α), the noise in the graph (σ, first plots) and the number of points (*n*, second plots). The third plots show that subjects responses depended not only on the prescribed slope α, but on the actual slope $\hat{\alpha}$ after the addition of noise. The fourth plots show that all the effects of noise, number of points and slope could be subsumed by an influence of the *t* value associated with the Pearson coefficient of correlation, as if participants performed a mental linear regression. In this graph, each dot represents the mean, across trials and subjects, of all data for one of the 112 experimental conditions determined by each combination of α, σ and *n*. Color figures are available in the online version of the article.

particular trial. The noise terms ε_I were generated independently for each trial and each participant. Due to this noise term, the actual linear regression line, as calculated from each dataset, could pass slightly above or below the center of the screen. To sidestep this issue, the coordinates were adjusted vertically by subtraction of the mean to ensure that the underlying regression actually passed through the exact center of the screen (i.e., through the point P having coordinates x = 0.5 and y = 0.5). The x and y coordinates were then rescaled to the coordinates of the computer screen used in the experiment, and each data point was represented by a 2-mm white dot centered at the appropriate location (~0.23° of visual angle given the distance of 50 cm from the screen). Fig. 1B shows four examples of scatterplots derived from datasets with different parameter values. As we can see, higher values of α correspond to higher inclinations of the graph, and higher values of σ result in noisier scatterplots.

2.1.5. Data availability

The data for this and following experiments is available on the Open Science Framework at: https://osf.io/z9qsw/.

2.2. Results

2.2.1. Statistics

The data (for all experiments) were primarily analyzed with repeated-measures ANOVAs. Whenever the assumption of sphericity was not met, we applied the conservative Greenhouse-Geisser correction, which corrects the overestimation of degrees of freedom that would result from a lack of sphericity (Greenhouse & Geisser, 1959).

2.2.2. Performance

We first looked at the proportion of "increasing" responses as a function of the prescribed slope α and either the prescribed noise σ or the number of points (Fig. 2, top). A repeated-measures ANOVA on the fraction of "increasing" responses confirmed a main effect of the prescribed slope ($F[6, 54] = 466.51$, partial $\eta^2 = 0.98$, $p < .0001$), and its interaction with noise ($F[18,162] = 12.62$, partial $\eta^2 = 0.58$, $p < .0001$) and with the number of points ($F[18,162] = 4.97$, partial $\eta^2 = 0.36$, $p < .0001$). In other words, the smaller the slope, the higher the influence of noise and number of points on the trend detection task.

We also conducted an analysis of each participant's sensitivity to prescribed slope, as a function of the noise σ and number of points *n*. We fitted a logistic regression to the percentage of "increasing" responses as a function of the prescribed slope, separately for each participant, noise level, and number of points, and used the slope of that logistic function as an indicator of sensitivity. In a few cases (11 out of 160 combinations), the logistic regression was not a meaningful indicator of performance since the data were better modeled through a step function (meaning that perfect answers were always given, thus resulting in a slope approaching to infinity); these cases were substituted with the maximum observed value. We submitted the resulting sensitivity values to a repeated-measures omnibus ANOVA and found a significant main effect of noise ($F[1.61, 14.45] = 82.72$, partial $\eta^2 = 0.90$, $p < .0001$) and a significant main effect of the number of points ($F[2.29, 20.60] = 19.33$, partial $\eta^2 = 0.68$, $p < .0001$). The direction and monotonicity of these effects indicated that, as expected, participants' decisions became increasingly sensitive as the datasets had a smaller noise level and a higher number of data points.

Because of the randomness in the stimulus generation process, on any given trial the actual slope of the linear regression line $\widehat{\alpha}$ could differ from its prescribed value α. We wondered if participants were sensitive enough to detect such variations in the actual slope of the graph. To this end, we looked at the fraction of "increasing" responses as a function of the prescribed slope α, while separating the trials based on whether the actual slope $\widehat{\alpha}$ was above or below α (Fig. 2 top, third plot from the left). A repeated measures omnibus ANOVA revealed a significant effect of the prescribed slope ($F[2.83, 25.51] = 549.13$, partial $\eta^2 = 0.98$, $p < .0001$), of the direction of the actual slope ($F[1, 9] = 3109.05$, partial $\eta^2 > 0.99$, $p < .0001$) and an interaction of the two factors ($F[2.37, 21.32] = 39.62$, partial $\eta^2 = 0.81$, $p < .0001$), meaning that participants were strongly influenced by the actual slope $\widehat{\alpha}$. We confirmed this effect by focusing on scatterplots with a prescribed slope of zero, and with an actual slope extremely close to zero (between −0.03 and + 0.03), and examined if participants were still able to extract the correct trend for those very hard trials. Indeed, the number of "increasing" responses was significantly larger for positive slopes (95/156 = 61%) than for negative slopes (68/170 = 40%; $\chi^2 = 14.21$, $df = 1$, $p < .001$), indicating a significant sensitivity even within this limited range.

The results so far indicate that participants' judgements were highly sensitive to the slope, the noise, and the number of points in a graph. We next tested the prediction of the "mental regression" hypothesis, according to which all of these effects may be subsumed by a single equation, the *t* value that a statistician would compute to judge whether a significant trend is present in the data. For each graph, we computed the Student *t* value associated to its Pearson coefficient of correlation and replotted the percentage of "increasing" responses as a function of that *t* value (Fig. 2, top right). As we can see, participants' mean performance was a sigmoid function of *t*. We compared the logistic regression of participants' responses as a function of either the actual slope ($\widehat{\alpha}$) or the *t* value. A simple model comparison based on the Akaike Information Criterion (AIC) values revealed that participants' responses were significantly better predicted by the *t* value (AIC for actual slope as predictor: 4138; AIC for *t* value as predictor: 4086.7; $\Delta_{AIC} = 51.3$, $p < 10^{-16}$). Furthermore, we replicated the above sensitivity analysis once the data were accounted for by the *t* value, and verified that the sensitivity values, once computed as a logistic function of t, were no longer affected either by σ or by *n* (respectively $F[1.4, 12.63] = 1.27$, partial $\eta^2 = 0.12$, $p = .3$ and $F[1.61, 14.53] = 1.18$, partial $\eta^2 = 0.12$, $p = .32$). In other words, the entire behavior was captured by a single value, the *t* value (Fig. 2 top, fourth plot from the left).

One might argue that the Pearson coefficient of correlation *r* may also provide a good model of human behavior. Indeed, *r* is a measure of the strength of a linear trend that jointly summarizes the effects of the slope α and the noise σ – but crucially, not the

number of points *n*. To investigate whether *r* alone sufficed to account for behavior, we plotted participants' percentage of "increasing" answers (averaged across conditions and subjects) either as a function of Pearson *r*, or as a function of *t* value (see supplementary figure S1). The graphs made it clear that the number of points *n* continued to play a significant role on participants' responses, over and above the effect of *r* alone, and that this effect was entirely captured by the *t* value. To put this observation to a test, we performed the same sensitivity analysis as above, as a logistic curve either as a function of *r* or of *t*. Sensitivity values, when computed as a logistic function of *r*, were still significantly affected by the number of points ($F[2.49, 22.39] = 10.64$, partial $\eta^2 = 0.54$, $p < .001$), whereas when computed as a logistic function of the *t* value, as already noted, the effect of number of points disappeared.

2.2.3. Response times

We then looked at the response times as a function of the prescribed slope and either the prescribed noise (Fig. 3, left) or the number of points (Fig. 3, middle). We conducted a repeated-measures omnibus ANOVA with median response times per condition as dependent variable, and prescribed slope, noise and number of points as within-participants factors. We found a main effect of slope ($F[6, 54] = 21.06$, partial $\eta^2 = 0.7$, $p < .0001$), a main effect of noise ($F[3, 27] = 20.07$, partial $\eta^2 = 0.69$, $p < .0001$), and an interaction of noise and slope ($F[18, 162] = 3.05$, partial $\eta^2 = 0.25$, $p < .0001$). As we can see from Fig. 3 (left), higher slope values and smaller noise values led to faster responses. As concerns the number of points, although there was no main effect of this variable ($F[3, 27] = 0.47$, partial $\eta^2 = 0.05$, $p = .7$), it entered into significant interactions with both slope ($F[18,162] = 1.88$, partial $\eta^2 = 0.17$, $p = .02$) and noise ($F[9,81] = 2.7$, partial $\eta^2 = 0.23$, $p < .01$). Once again, as shown in Fig. 3 (right), those effects on median response times were well summarized by a single function of the *t* value associated to the regression: RTs varied as a symmetrical, convex-upward, monotonously decreasing function of the distance of the *t* value from zero.

Could the shape of this RT effect be predicted by the mental Pearson model? Following Gold and Shadlen (2002), we assumed that participants based their decisions on a noisy accumulation of evidence towards a fixed decision bound. Given our theoretical assumptions (see Appendix), we assumed that the noisy samples upon which the decision was based arose from a sampling of the regression *t* value. Under those assumptions, the probability P of responding "increasing" and the mean RTs should follow the following joint equations (see Gold & Shadlen, 2002):

$$P(t) = \frac{1}{1 + e^{-2Bt}}$$

and

$$RT(t) = \frac{B}{|t|}\tanh(B|t|)$$

where B is a free parameter that corresponds to the slope of the psychometric function. We first fitted B using the performance data presented in Fig. 2 (top right), then plugged this value into the second equation to obtain the shape of the predicted RT as a function of the *t* value in our stimuli. We thus obtained dimensionless predicted RTs for each experimental condition, which we fitted to the data using a 2-parameter linear regression where the dependent variable was the across-participants mean RT in each of the 112 experimental conditions. The model provided a very good fit to the participants' RTs ($r^2 = 0.76$; regression slope = 286 ± 15 ms, $t(1\ 1\ 0) = 18.97$, $p < 10^{-16}$; intercept (non-decision time) = 429 ± 11 ms, $t(1\ 1\ 0) = 37.43$, $p < 10^{-16}$). The corresponding curve is shown in blue in Fig. 3 (right).

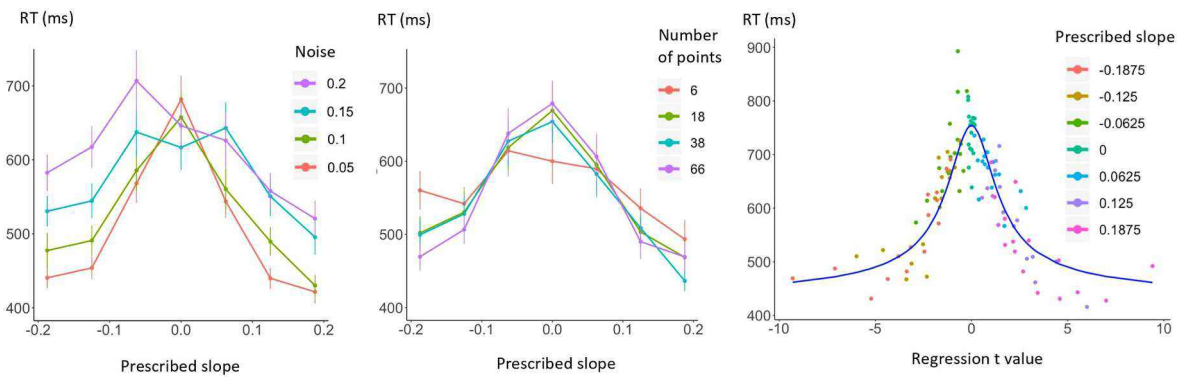

Fig. 3. Response times in experiment 1 (trend judgment). Mean response times are shown as a function of the prescribed slope (α) and either the noise (σ, left) or the number of points (*n*, middle). Error bars indicate one standard error of the mean across subjects. The plot on the right shows the response times as a function of the mean *t* test value associated to the Pearson coefficient of correlation, within each of the 112 cells of the experimental design. The blue line indicates the response times predicted by a simple accumulation-of-evidence model (Gold & Shadlen, 2002). Color figures are available in the online version of the article. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.3. Discussion

The results from the first experiment reveal that participants are able to quickly extract the linear trend of a scatterplot, without requiring any sophisticated training or long exposures to the stimulus. The fast presentation time (100 ms) and the short response times we observed (below 900 ms on average, see Fig. 3) imply that participants did not have time to perform complex calculations. Rather, they must have relied on an intuitive yet accurate estimate of the correlation. In fact, even on trials with a prescribed slope equal to 0, participants' performance remained above chance level, indicating a fine sensitivity to random variations in the graphs.

As expected, all three parameters of slope, noise and number of points significantly affected participants' accuracy, making it lower for shallower slopes, higher noise levels and smaller datasets. Similar effects were observed on RTs, and all those effects were subsumed by the *t* value predicted by Pearson coefficient of correlation. By applying the model of Gold and Shadlen (2002) to our data, we found that participants' decisions followed the prediction of a classic accumulation-of-evidence decision model, where the decision variable was the strength of the *t* value associated with the Pearson correlation coefficient. This finding suggests that, before giving an answer, participants were accumulating evidence on the dataset's trend, and that this decision process approximated a statistical regression procedure. Indeed, participants' performance was better modeled as a function of the *t* value rather than as a function of the prescribed slope. Thus, when detecting a scatterplot's tendency, human adults do not solely rely on the slope of the linear regression, but extract an approximate summary statistic.

It is noteworthy that the mean RTs did not increase with the number of points *n*. On the contrary, RTs either stayed roughly constant or even decreased with *n* for large values of the slope α (see Fig. 3). Thus, participants did not treat the data points serially, as would have been unavoidable if the data were presented through numbers (such as in a tabular form), but took advantage of the graphic presentation to process them in parallel. We conclude that the human visual system affords a parallel form of approximate regression. Note that the coefficient of correlation formula involves only variances (in x and y) and covariances, all of which are sums or, equivalently, averages over values provided by each data point – and there is considerable prior evidence that the visual system can compute averages of various features in parallel across the items in a set ("ensemble perception"; see e.g. Chong & Treisman, 2003, 2005; Van Opstal, de Lange, & Dehaene, 2011). The present work extends this concept to the case of statistical trend perception.

The present results go beyond previous studies (Cleveland et al., 1982; Lane, Anderson, & Kellam, 1985; Rensink & Baldridge, 2010) which showed that, when participants are asked to judge the strength of an association presented in a scatterplot, their judgement is more accurate for higher levels of the Pearson correlation coefficient *r*, although still affected by the variance of the dataset (Lane et al., 1985). As pointed out by Surber (Surber, 1986), however, the interpretation of subjective ratings of correlation is

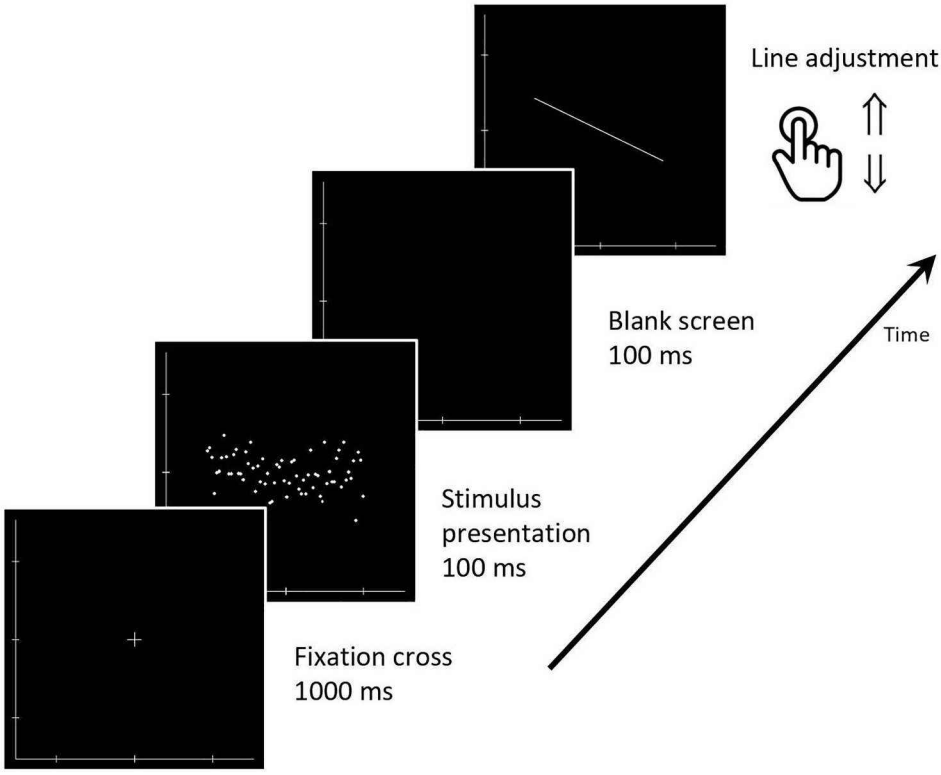

Fig. 4. Experimental design for experiment 2 (line fitting). Subjects were presented with a simple scatterplot, generated by a linear function plus noise, with a variable number of data points. Immediately after, they were asked to adjust a line by moving their finger on a trackpad, in order to provide an estimation of the regression line underlying the noisy scatterplot.

complex and may be hard to relate to objective statistics. In agreement with this observation, Cleveland and colleagues (Cleveland et al., 1982) found that the perceived correlation was particularly small for values of $r < 0.5$, suggesting that only for high levels of r (>0.95) participants' perceived correlations might relate to the actual r value. For these reasons, in our study, we decided to avoid subjective measures of perceived correlation and asked for a more direct categorical trend judgement (increasing or decreasing). This forced-choice task provided solid evidence that participants could perform a "mental regression" with a high sensitivity, even when the slope of the linear trend in the data was shallow and the noise was high.

3. Experiment 2: Slope estimation

In experiment 2, we further probed whether human participants act as good statisticians, and which procedure they use to approximate linear regression. Whereas experiment 1 solely asked participants to decide whether an ascending or descending trend was present, we now asked participants to report the slope of the best-fitting regression line. They did so by using a trackpad to adjust the tilt of a line on screen.

What predictions can we derive for this task? If participants used a mental process equivalent to simple linear regression (also called "ordinary least squares" method, OLS), then their average slope estimate should be centered on the prescribed slope used to generate the graph, and should not be influenced by the number of data points (n) nor by the noise level (σ). The reason is that the OLS estimate of slope is what statisticians call an "unbiased estimator", i.e. an estimate whose expected value is equal to the prescribed slope (McElroy, 1967; Puntanen & Styan, 1989). In experiment 2, we tested whether participants' slope estimates follow this law.

3.1. Methods

3.1.1. Participants

10 participants were recruited for the experiment (age: 25.2 ± 1.2, 4 females, 6 males). All participants had to meet the same inclusion criteria of experiment 1. They were paid 10 euros for their participation. The experimental session lasted approximately 45 min. The experimental procedure was approved by the local ethical committee. One participant was excluded by the analyses because he failed to perform the task appropriately (he did not adjust the line for more than half of trials).

3.1.2. Procedure

Stimuli and procedure were identical to experiment 1, except that immediately after the scatterplot (presented for 100 ms), a blank screen appeared for 100 ms and then an adjustable line was shown in the middle of the screen (see Fig. 4). The line was initially horizontal, but participants were asked to adjust it as accurately as possible by moving their right index on the computer trackpad. The center of the line was kept fixed (since, as in experiment 1, the OLS regression line of the scatterplot always passed through the exact center of the graph), so that moving the finger up or down the trackpad resulted in a rotation of the line around its center, whose angle was proportional to finger displacement; moving the finger up tilted the line in the counterclockwise direction, whereas moving the finger down tilted it in the clockwise direction. The participants were informed that we would measure the accuracy of their fit and, for this reason, they were invited to take their time to perform the task. When the adjustment was completed, they simply had to press the trackpad in order to confirm their answer and move to the next trial, which was, as in experiment 1, preceded by a 1 s fixation cross.

Exactly as for experiment 1, the task was divided into 6 blocks, each comprising all 112 conditions; the duration of each block was now ~ 6 min. After each block, the participants could take a short break and received feedback on the total number of correct responses they gave (for this feedback, a response was considered as correct if its slope had the correct sign, i.e. positive or negative). Before the beginning of the actual experiment, 25 practice trials were conducted under the supervision of the researcher, in order to control for the correct execution of the task.

3.2. Results

3.2.1. Increasing or decreasing judgments: Replication of experiment 1

To see if we could replicate the findings of experiment 1, we first categorized the participants' responses as increasing or decreasing (based on the slope of their regression lines) and examined the proportion of "increasing" responses as a function of the prescribed slope and either the prescribed noise (Fig. 2, bottom left) or the number of points (Fig. 2, bottom middle). A repeated measures ANOVA on the percentage of "increasing" responses confirmed the statistical significance of the data presented in the figures, which closely paralleled the findings from experiment 1: we again found a main effect of the prescribed slope ($F[6, 48] = 508.51$, partial $\eta^2 = 0.98$, $p < .0001$) and its interaction with noise level ($F[18,144] = 12.36$, partial $\eta^2 = 0.61$, $p < .0001$) and number of points ($F[18,144] = 8.88$, partial $\eta^2 = 0.53$, $p < .0001$). Once again, the closer the slope was to zero, the higher the influence of the noise and of the number of points (Fig. 2). As in experiment 1, we conducted an analysis of each participant's sensitivity as a function of the noise σ and number of points n. We fitted a logistic regression to the percentage of "increasing" responses (as a function of the prescribed slope), separately for each participant, noise level, and number of points, and used the slope of the logistic function as an indicator of sensitivity. In the few cases (5 out of 144 combinations) where the logistic regression was not meaningful, since the data were better modeled by a step function, the values were substituted with the maximum observed value. We submitted the resulting sensitivity values to a repeated-measures omnibus ANOVA and found a significant main effect of noise ($F[3, 24] = 96.65$, partial $\eta^2 = 0.92$, $p < .0001$) and a significant main effect of the number of points ($F[3, 24] = 24.45$, partial $\eta^2 = 0.75$, $p < .0001$). As we can see from Fig. 2 (bottom left and middle), participants were significantly more sensitive to datasets having a smaller noise and a higher number of points, closely replicating the

results from experiment 1. We also looked again at the fraction of "increasing" responses as a function of the prescribed slope and the direction of the actual slope compared to the prescribed one (i.e., above or below it): a repeated-measures omnibus ANOVA revealed a significant effect of the prescribed slope ($F[2.03, 22.44] = 563.86$, partial $\eta^2 = 0.99$, $p < .0001$), of the direction of the actual slope ($F[1, 8] = 437.65$, partial $\eta^2 = 0.98$, $p < .0001$) and an interaction of the two factors ($F[2.74, 21.96] = 29.79$, partial $\eta^2 = 0.79$, $p < .0001$), meaning that participants, once again, were able to base their judgement on the actual value $\widehat{\alpha}$ rather than the prescribed slope α.

Exactly as for experiment 1, for each graph, we computed the Student *t* value associated to its Pearson coefficient of correlation and

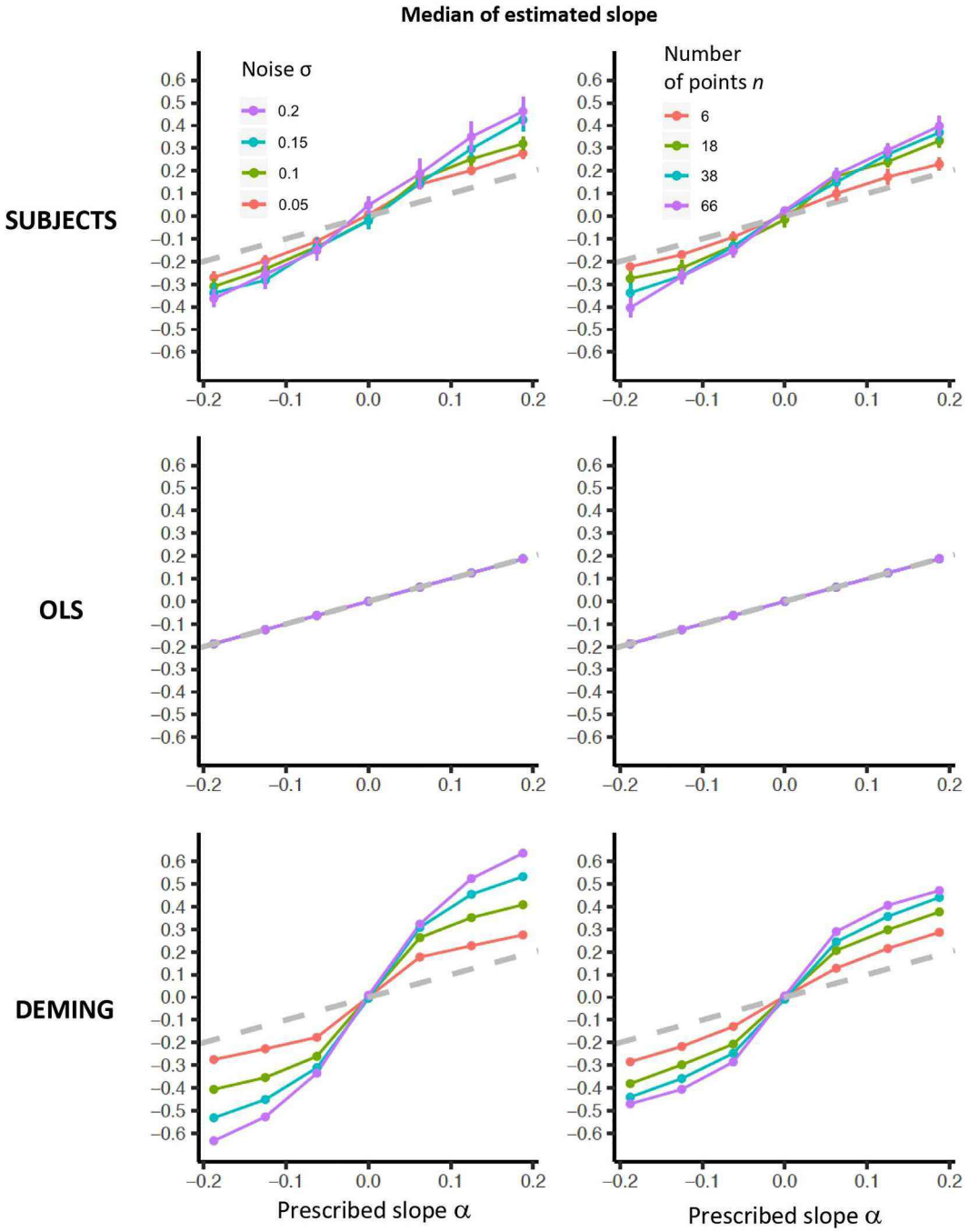

Fig. 5. Slopes reported by the participants (top) and predicted by the OLS and Deming regression models (middle and bottom). Values are plotted as a function of the prescribed slope α, noise σ and number of points *n*. The dashed line represents the ground truth, i.e. the prescribed slope α. Subjects' median estimated slopes show a bias similar to Deming predictions. Color figures are available in the online version of the article.

replotted the percentage of "increasing" responses as a function of that *t* value (Fig. 2, bottom right). As we can see, participants' mean performance was once again a sigmoid function of *t*. We compared the logistic regression of participants' responses as a function of either the actual value of the slope ($\widehat{\alpha}$) or *t*. A simple model comparison based on the Akaike Information Criterion (AIC) values revealed that participants' responses were again significantly better predicted by the *t* value (AIC for actual slope as predictor: 4815; AIC for *t* value as predictor: 4574; $\Delta_{AIC} = 241$, $p < 10^{-16}$). We also replicated the above sensitivity analysis once the data were accounted for by the *t* value, and verified that neither σ nor n played a significant role (respectively: $F[3, 24] = 2.69$, partial $\eta^2 = 0.25$, $p = .07$ and $F[3, 24] = 1.04$, partial $\eta^2 = 0.12$, $p = .39$), confirming the results from experiment 1: the entire behavior was captured by a single value, the *t* value (Fig. 2 bottom, fourth plot from the left).

3.2.2. Slope estimation

Fig. 5 (top) shows the slope estimates, averaged across participants, as a function of the prescribed slope and either the noise level (left) or the number of points (right). We conducted a repeated measures ANOVA on participants' median estimated slopes. As expected, we found a main effect of slope ($F[6, 48] = 91.6$, partial $\eta^2 = 0.92$, $p < 0.001$): as the prescribed slope increased continuously across 7 levels, so did the participants' estimates. However, the values that they reported were always in excess of the ideal slopes, both in the positive and in the negative direction (see Fig. 5, dashed line). Furthermore, this tendency to exaggerate the linear trends increased with noise level, and also with the number of points, as attested by significant interactions of prescribed slope and noise level ($F[18, 144] = 3.56$, partial $\eta^2 = 0.31$, $p < 0.001$), and prescribed slope and number of points ($F[18, 144] = 9.63$, partial $\eta^2 = 0.55$, $p < 0.001$), as well as a triple interaction of slope, number of points and noise ($F[54, 432] = 2.07$, partial $\eta^2 = 0.21$, $p < 0.001$). The nature of this bias can be described as follows. First, participants always overestimated positive slopes, and did so with a bias that increases with noise level and number of points (ANOVA restricted to positive slopes: main effect of noise, $F[3,24] = 6.43$, partial $\eta^2 = 0.45$, $p < 0.01$; main effect of number of points, $F[3,24] = 12.48$, partial $\eta^2 = 0.61$, $p < .0001$). Second, conversely, participants always underestimated negative slopes, again increasingly so for larger noise levels and numbers of points (ANOVA restricted to negative slopes: main effect of noise, $F[3,24] = 3.48$, partial $\eta^2 = 0.30$, $p = 0.03$; main effect of number of points, $F[3,24] = 20.59$, partial $\eta^2 = 0.72$, $p < .0001$).

3.3. Discussion

The significant dependency of participants' estimated slope on both noise and number of points violates the predictions of simple linear regression (OLS). Since OLS slopes are unbiased statistical estimators of the true underlying slope, OLS predicted no effect of either noise or number of points on the slope estimates (Fig. 5, middle). Those predictions were clearly violated in the data. Note in particular that the more data points were present, the more the participants' slope estimates were biased towards exceedingly extreme values. This finding may seem paradoxical, given that in OLS, a larger number of data points implies that the regression can be estimated with greater precision – and such an effect was indeed found in participants' proportion of "increasing" responses for a fixed prescribed slope, in both experiments 1 and 2 (Fig. 2).

How can we explain the participants' behavior? A key observation is that simple linear regression, based on ordinary least squares,

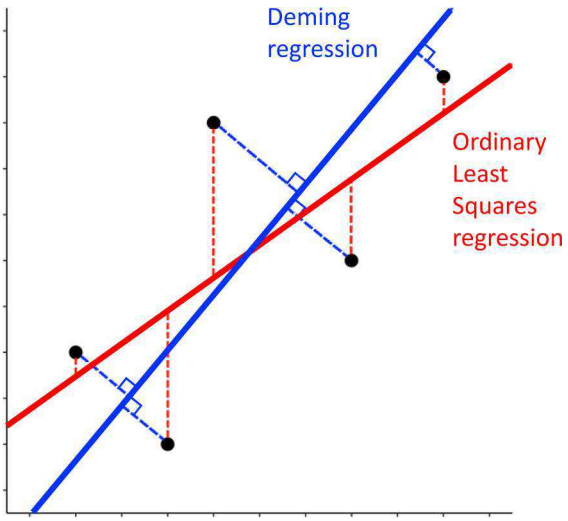

Fig. 6. Illustration of the difference between ordinary least squares (OLS) and Deming regression. OLS regression (red line) minimizes the sum of the squares of the vertical distances of the points to the line. It is appropriate when the x values are fixed and there is noise only in the dependent variable (y axis). Deming regression (blue line) minimizes the sum of the squares of the orthogonal distances of the points to the line (assuming equal variance on the x and y measures). It is appropriate when the measurement is noisy on both the x and y axes. Color figures are available in the online version of the article. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

is not the only procedure that can be used to estimate the slope of a graph. Indeed, it is not even always the optimal one. If there is variance in both the x and the y measurements, statisticians recommend the use of another procedure termed "Deming regression" (Deming, 1943; Linnet, 1998; Martin, 2000). This procedure belongs to the category of "errors-in-variables models" which optimally account for the fact that there can be measurement errors on both x and y axes. The gist of Deming regression is illustrated in Fig. 6 (blue line). Essentially, it can be conceived as an "orthogonal regression", seeking the line that minimizes the sum of square distances to the data *simultaneously in both the x and y dimensions* (strictly speaking, this is true only when assuming equal noise on x and y; otherwise one of the axes must first be scaled). This is the appropriate thing to do if x itself is the result of a noisy measurement. Classical regression (OLS), on the other hand (Fig. 6, red line), minimizes the sum of square distances only along the y axis; this is the proper thing to do if x is a fixed experimental factor, and only y is a noisy measure.

OLS regression is clearly appropriate for our graphs, which were created by keeping x fixed and generating y using a linear equation plus noise. However, participants, unaware of this fact, might have applied a procedure akin to Deming regression, perhaps because they treated the x and y coordinates of the dots as equivalent and therefore both potentially subject to noise. Indeed, Deming regression presents the unique advantage of yielding an identical regression line whether y is regressed on x, or x is regressed on y. This is not true of OLS regression, which treats x and y asymmetrically. Note, however, that the correlation coefficient r, and therefore the t test, are symmetrical in x and y, and are appropriate measures of the presence and strength of a linear trend for both OLS and Deming regression.

Fig. 5 (bottom) shows the slopes predicted by Deming regression (calculated over 50,000 stimuli generated through the identical algorithm used to generate the stimuli presented to participants). Remarkably, the Deming model made predictions strictly parallel to what we observed in our data: the median slope increased with both the noise level and with the number of data points. For Deming regression, one must provide not only the x and y values, but also the ratio of the variances of the errors on x and on y. Here, we used the ratio of the empirical variances in the graphs, but qualitatively similar predictions were obtained if we assumed a fixed ratio of 1.

To summarize, experiment 2, using a line adjustment procedure, closely replicated the results of experiment 1 with binary trend judgement (increasing or decreasing). This parallelism suggests that, when asked to quickly extract the tendency of a scatterplot through a simple binary choice, humans can be as accurate as when precisely adjusting a regression line. Remarkably, both fast (experiment 1) and slow (experiment 2) judgments were affected by the same stimulus parameters, namely the slope, the noise and the number of points. Crucially, behavior was again subsumed by the t value associated to the Pearson coefficient of correlation. This finding offers a methodological guidance for future experiments in graph perception: fast binary choices might be as informative as slow line adjustments when investigating the human perception of positive or negative trends in scatterplots.

However, slope adjustment also revealed a result that was inaccessible to binary judgments: humans are biased in their estimations of linear trends. They overestimate positive slopes, underestimate negative ones, and those biases increase with noise and with number of points. These findings refute the hypothesis that human adults compute a traditional OLS regression, and instead suggest that participants might use Deming regression when fitting a line to a noisy scatterplot. Simulations showed that Deming regression, far from being unbiased, leads to exactly the same qualitative biases as observed in humans.

Deming regression feels reasonable because it essentially consists in finding a line that minimizes the Euclidean distances to all points, thus treating the cloud of dots as a 2-dimensional shape, without distinguishing the x and y measurements (as OLS does). Thus, Deming regression, yields the same line whether y is regressed on x or vice-versa, unlike OLS. Deming regression might have been induced by the stimuli we used, which were square graphs with identical layouts for the x and y axes, thus perhaps encouraging participants to treat the x and y axes as two noisy measurements. However, note that the x values were always equally spaced discrete samples, a fact that was particularly obvious for small numbers of points (see Fig. 1, $n = 6$); yet even in this case, the Deming-like bias was present. Thus, our findings suggest that human participants fail to apply the most standard regression procedure, ordinary least squares, and exhibit a strong bias, whose consequences will be reexamined in the general discussion.

4. Experiment 3: Linear extrapolation

Experiments 1 and 2 indicate that human adults can categorize a linear trend as ascending or descending, and approximate its slope. In experiment 3, we examined if their intuitive statistical skills also allowed them to perform a third task: linear extrapolation. We refer to extrapolation as an estimation that is made beyond the original observation range, assuming that the trend underlying the scatterplot will continue to be the same. To test it, we engaged participants in an extrapolation task, in which they adjusted a point vertically to place it on their best estimation of the regression line fitting the data points. This instruction aimed to minimize the tendency of participants to add noise to their extrapolations in order to match the noise in the graph, a phenomenon already described by Bolger and Harvey (1993).

Our predictions were simple: if participants relied on Deming regression, then they should produce exaggerated estimates (deviating too far either upwards or downwards, depending on whether the main slope is positive or negative), and this bias should be all the more pronounced that the noise in the scatterplot is high.

4.1. Methods

4.1.1. Participants

10 participants were recruited for the experiment (age: 24.6 ± 1.8, 5 females, 5 males). All participants met the same inclusion criteria as in experiment 1. They were paid 5 euros for their participation. The experiment lasted approximately 30 min and was approved by the local ethical committee.

4.1.2. Experimental design and stimuli

The stimuli were generated according to the same algorithm of experiment 1 and 2. Five levels of prescribed slopes were used to generate the scatterplots (−0.48, −0.24, 0, +0.24, +0.48). The number of points was kept fixed at 18. The noise levels were the same used in experiments 1 and 2 (0.05, 0.1, 0.15, 0.2). The scatterplot was now confined to the left part of the screen, while the right one served as the extrapolation area (see Fig. 7A). The location of the scatterplot was vertically jittered by a random amount in order to induce participants to avoid responding at the same location; the jitter was later corrected for in our analyses. We included a considerable margin (12.5% of the screen) above and below the locations of the correct answers, where no expected correct answers could fall into; this was done in order to allow participants to give a free and unconstrained response, even if considerably higher or lower than the correct one.

4.1.3. Procedure

The procedure closely followed experiments 1 and 2, except that each scatterplot was now presented on the left side of the screen and for a long duration (until the response). On the right side, a single point was shown at one of two possible x coordinates (either x = 1.3 or x = 1.6, which we refer to as "probed positions") and at a y coordinate corresponding to the middle of the y axis (Fig. 7A). Participants were asked to vertically adjust the point as accurately as possible by moving their right index on the computer trackpad.

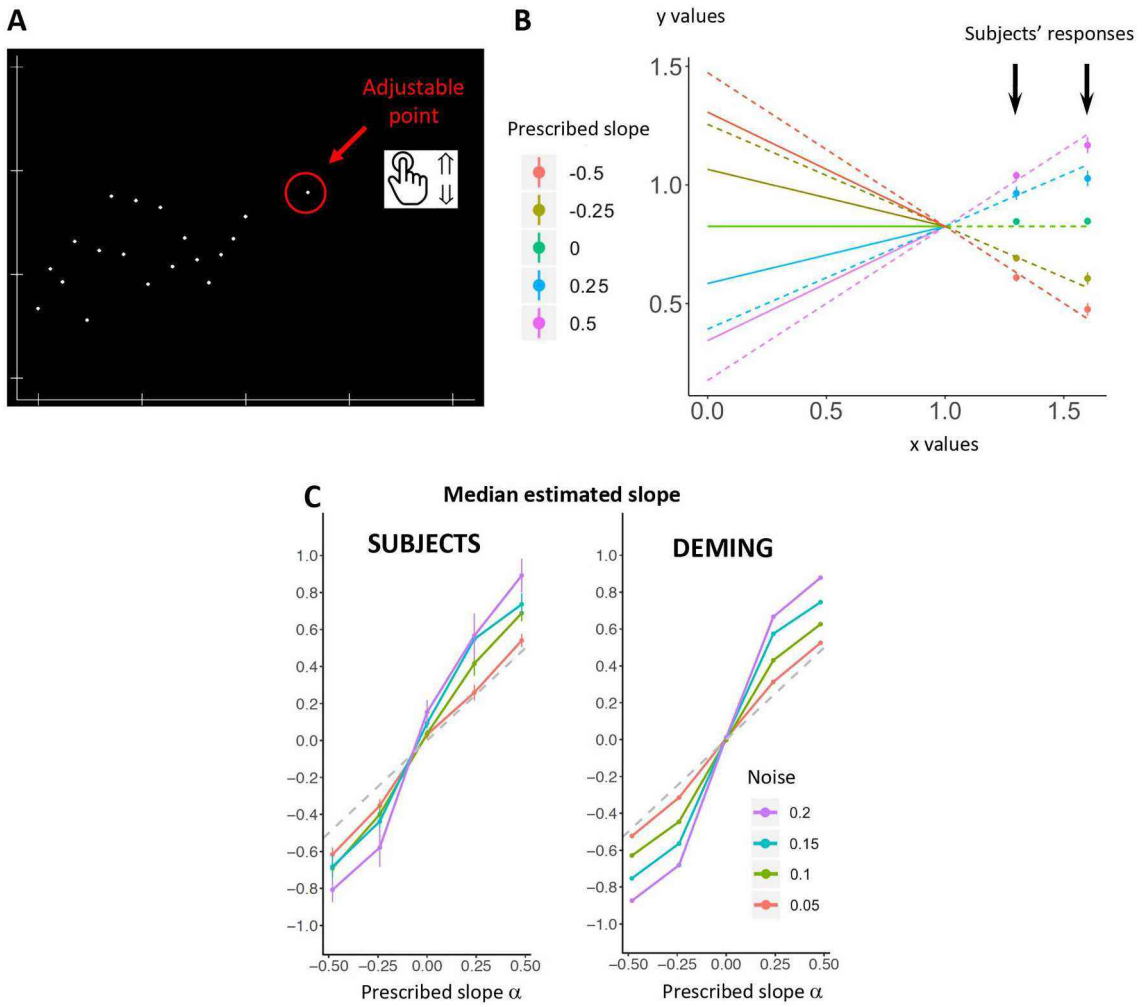

Fig. 7. Results of experiment 3 (linear extrapolation). A. Left: example of trial. Participants adjusted a movable dot (here indicated by a red circle and a red arrow, for illustrative purposes only) in the vertical direction so that it best extrapolated the data at left. B. Subjects' median extrapolation responses at x = 1.3 and x = 1.6. The solid lines indicate the OLS regression lines, which match the original lines from which the noisy scatterplots were generated. The dashed lines indicate the median slopes predicted by Deming regression. C. Mean (and standard error) of the median estimated slope computed from the subjects' responses (left), and compared to the median slope predicted by Deming regression (right). The slope was calculated as the slope of a line passing through the center of the screen and through the answer given by the subjects. The results in the plot are averaged across the two probed x positions. Color figures are available in the online version of the article. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Once they were satisfied with the given response, they confirmed it by pressing the trackpad. Participants were explicitly asked to give an intuitive answer and to locate the point on their best estimation of the regression line of the scatterplot. The task was divided into 7 blocks, each comprising one trial for each of the 40 conditions (5 slopes, 4 noise levels, two probed positions), for a total of 280 trials. The duration of each block was ~ 4 min. After each block, the participants could take a short break. No feedback was given. Before the beginning of the actual experiment, 25 practice trials were conducted under the supervision of the researcher, in order to control for the correct execution of the task.

4.2. Results

4.2.1. Location of the extrapolated point

For each of the 40 conditions and each participant, we computed the median response location on the y axis. We conducted a repeated measures ANOVA on those median extrapolation values with prescribed slope, noise and probed position as within-participants factors, and we found a significant effect of the slope ($F(4, 36) = 95.15$, partial $\eta^2 = 0.91$, $p < .0001$); no main effect of the probed position was found ($F(1,9) = 0.01$, partial $\eta^2 = 0.001$, $p = .92$), although it entered into a significant interaction with the slope ($F(4, 36) = 36.30$, Partial $\eta^2 = 0.80$, $p < .0001$). As shown in Fig. 7B, those findings reflect the fact that participants adapted their extrapolation responses to the prescribed slope and the probed position. However, as in experiment 2, an interaction of prescribed slope and noise ($F(12, 108) = 4.24$, partial $\eta^2 = 0.32$, $p < .0001$), as well as a triple interaction of slope, probed position and noise ($F(12, 108) = 2.46$, partial $\eta^2 = 0.21$, $p < .01$) indicated that, as the noise increased, the extrapolation responses became increasingly biased towards exaggerated values, as predicted by Deming regression. Specifically, as in experiment 2, participants always overestimated positive slopes, and did so with a bias that increases with noise level (ANOVA restricted to positive slopes: main effect of the noise, $F[1.68, 14.85] = 6.60$, partial $\eta^2 = 0.42$, $p = .01$). Conversely, participants always underestimated negative slopes, again increasingly so for larger noise levels (ANOVA restricted to negative slopes: main effect of noise level, $F[1.81,16.30] = 4.17$, partial $\eta^2 = 0.32$, $p = .04$).

Next, we directly compared participants' extrapolation responses with those expected under Deming regression. The solid lines in Fig. 7A show the functions from which the scatterplots were generated (i.e., the OLS regressions of the dataset), whereas the dashed lines show the Deming regressions. As already described above and in Figs. 5 and 6, Deming regression results in steeper slopes than OLS predictions. Relative to those lines, we can see that the participants' responses lay close to Deming predictions, although at $x = 1.6$ they tend to be slightly lower. To quantitatively test for the resemblance of participants' extrapolation responses to Deming predictions, we calculated the median slope associated with each extrapolated point (calculated as the slope of the line passing through the center and the given point). Fig. 7C shows the remarkable similarity of participants' extrapolations with Deming predictions. We conducted a repeated measures omnibus ANOVA on participants' median slopes and found a significant main effect of the prescribed slope ($F[4, 36] = 79.01$, partial $\eta^2 = 0.9$, $p < .0001$), a significant main effect of noise ($F[3, 27] = 3.17$, partial $\eta^2 = 0.26$, $p = .04$) and an interaction effect of the slope with the noise ($F[12, 108] = 4.36$, partial $\eta^2 = 0.33$, $p < 0.0001$). Although there was no main effect of the probed position ($F[1, 9] = 0.5$, partial $\eta^2 = 0.05$, $p = .5$), it entered into a significant interaction with the slope ($F[4,36] = 7.82$, partial $\eta^2 = 0.46$, $p < .0001$) and into a triple interaction with slope and noise ($F[12, 108] = 3.51$, partial $\eta^2 = 0.28$, $p < .001$), confirming that participants adapted their responses to the prescribed slope and the probed position. These effects reveal that the median slopes associated to the extrapolation responses were increasingly steeper as the noise increased, with a bias, once again, that increased with noise level (ANOVA restricted to positive slopes: main effect of noise, $F[1.64, 14.78] = 7.86$, partial $\eta^2 = 0.47$, $p < .01$; ANOVA restricted to negative slopes: main effect of noise, $F[1.67,15.07] = 6.22$, partial $\eta^2 = 0.41$, $p = .01$).

4.3. Discussion

The results of the third experiment showed that participants were able to perform an intuitive extrapolation, meaning they could predict the location of a point outside the range of available data. Unlike the predictions of ordinary least squares, participants' estimates were biased and were affected by noise level. In agreement with the results of experiment 2, their estimates resembled again those predicted by Deming regression.

We conclude that participants can approximate a linear regression from a noisy scatterplot, and do so with a comparable performance regardless of the details of the stimuli and the task: binary judgement on a flashed graph (experiment 1), slope adjustment on a flashed graph (experiment 2) or extrapolation on a long-exposure graph (experiment 3). The results of our three experiments converge to suggest that humans behave in a highly competent and consistent way when extracting statistical information from a scatterplot. They take into account not only the slope, but also the noise level and the number of data points – but do not do so according to classical OLS regression, but to the lesser known Deming regression.

5. Experiment 4: Extrapolation of non-linear curves

Linear regression is a standard procedure that, in most statistical packages, is typically applied without any verification of whether a linear fit does or does not make sense (e.g. if the data actually follows a non-linear trend). In experiment 4, we asked if humans are superior in the sense that they can recognize whether and when such a regression is appropriate. In experiment 4, using the same extrapolation task as in experiment 3, we extended the scatterplots to non-linear functions, and asked whether humans could adequately adapt their extrapolations to the specific function exemplified in the graph. Our hypothesis was that human adults may be able to go beyond linearity and spontaneously identify non-linear statistical trends, although their performance might still be affected

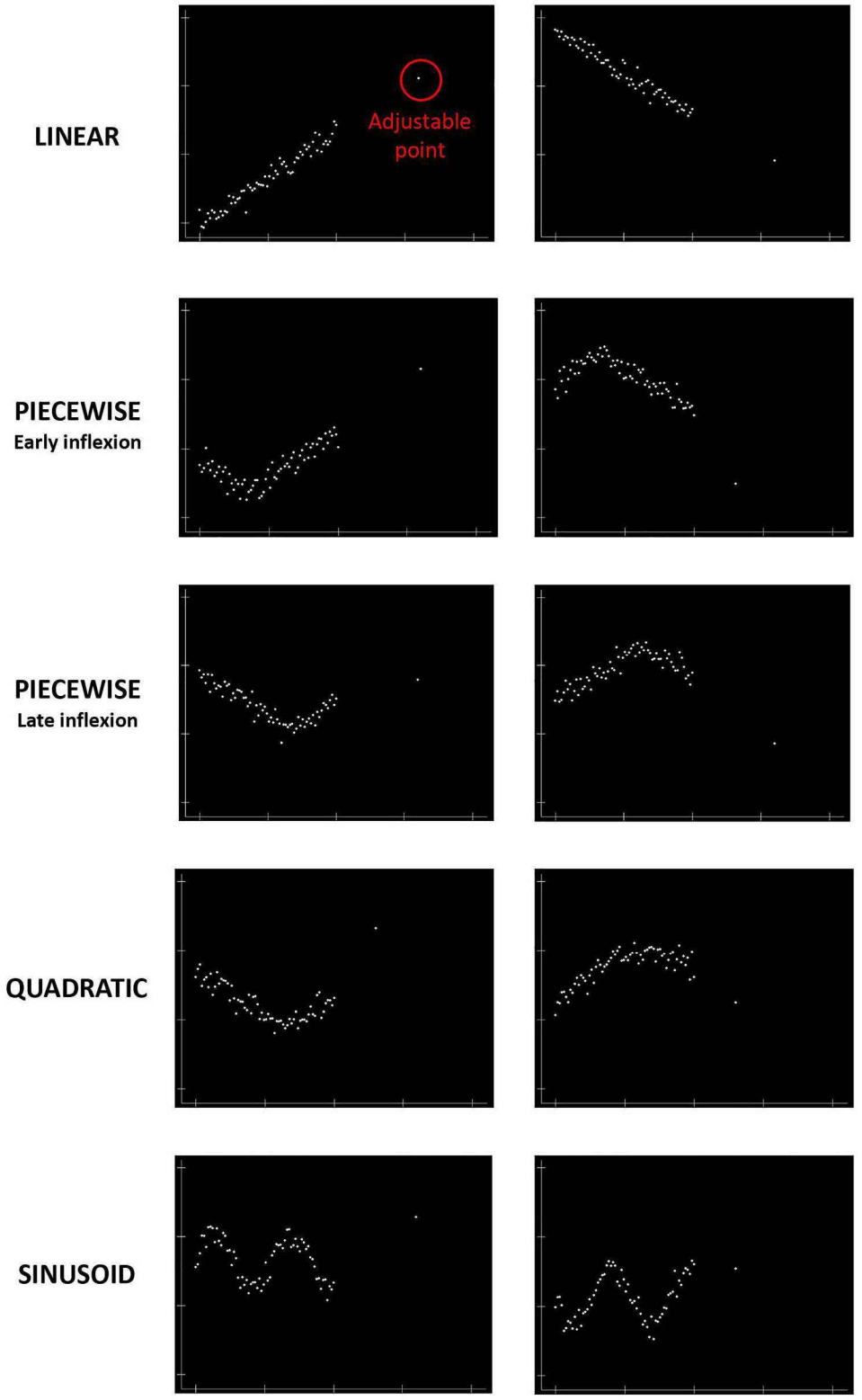

Fig. 8. Design of experiment 4 (linear and non-linear extrapolation). The figure shows examples of stimuli for each generative function used. The movable dot is indicated by a red circle. Color figures are available in the online version of the article. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

by a preference for linear trends (Kalish et al., 2007; Little & Shiffrin, 2009; Mcdaniel & Busemeyer, 2005). Specifically, our experiment was designed to disentangle two possible views of human extrapolation. Under hypothesis 1, participants would be restricted to linear extrapolation, based either on the last few points of a curve, or the tangent to the curve. Under hypothesis 2, participants would infer the nature of the curve and adapt their extrapolation correctly, either by taking into account the curvature in the last few points of the curve, or even the entire underlying function.

5.1. Methods

5.1.1. Participants

10 participants were recruited for the experiment (age: 23.1 ± 3, 5 females, 5 males), with the same inclusion criteria as in experiments 1, 2 and 3. They were paid 5 euros for their participation. The experiment lasted approximately 25 min and was approved by the local ethical committee.

5.1.2. Experimental design and stimuli

The stimuli were generated according to the same algorithm as experiments 1, 2 and 3, but five different functions were used to

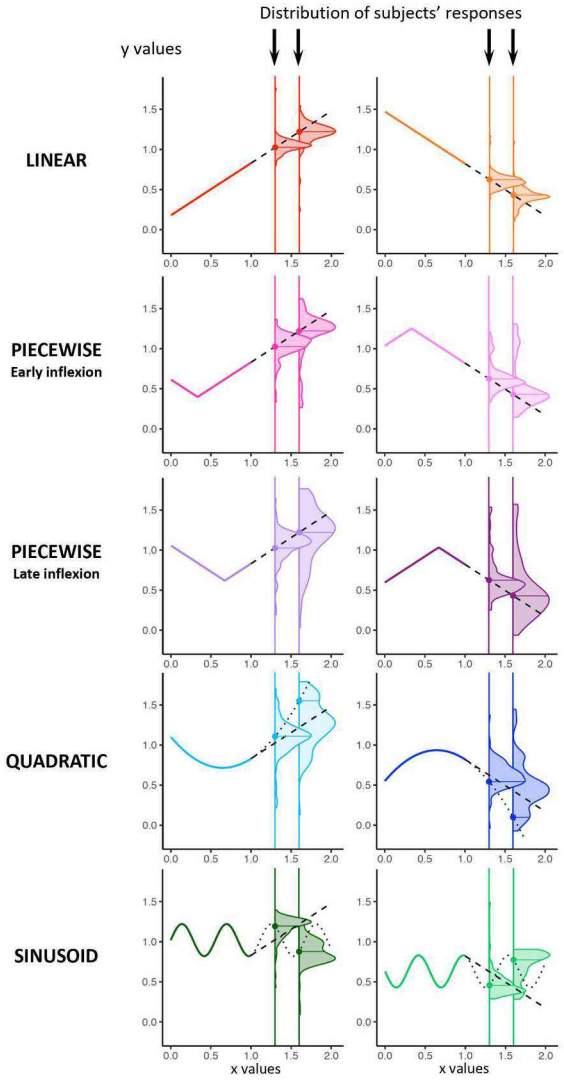

Fig. 9. Results of experiment 4 (linear and non-linear extrapolation). The functions from which the scatterplots were generated are shown on the left, and are prolonged by dotted lines. Large dots indicate the ideal answers at the two probed x positions (x = 1.3 and x = 1.6). Dashed lines show the linear extrapolation, based on the function's derivative at the last point. Density plots show the distribution of all given answers from all subjects for a given x position. For greater readability, all densities were normalized to the same peak height. Color figures are available in the online version of the article.

generate the scatterplot. All had the same absolute value of the derivative at their rightmost point within the interval plotted (x = 1), but half of them had a positive derivative (Fig. 8, left column) and half of them had a negative derivative (Fig. 8, right column). The functions were the following:

1) Two linear functions having equations $l_1(x) = 0.65x + 0.18$ and $l_2(x) = -0.65x + 1.47$.
2) Four piecewise linear functions composed of two straight-line segments. Two functions had their inflection point early on, at 1/3 of their length; equations: $pl_1(x) = 0.65\left|\frac{1}{3} - x\right| + 0.39$ and $pl_2(x) = -0.65\left|\frac{1}{3} - x\right| + 1.26$. The other two piecewise linear functions had a late inflection point, i.e. at 2/3 of their length; equations: $pl_3(x) = 0.65\left|\frac{2}{3} - x\right| + 0.61$ and $pl_4(x) = -0.65\left|\frac{2}{3} - x\right| + 1.04$.
3) Two quadratic functions, either convex or concave, with equations $q_1(x) = 0.92x^2 - 1.19x + 1.09$ and $q_2(x) = -0.92x^2 + 1.19x + 0.56$. Crucially, the x coordinate of, respectively, the minimum and maximum of these two functions (x = 0.65), was very similar to the x coordinate of the inflection point of the two piecewise linear functions with a late inflection (x = 0.66). These quadratic functions allowed us to examine if participants would correctly estimate the curvature of the graph, or even its entire quadratic trend.
4) Two sinusoidal functions completing 1.5 periods on screen and with two opposite phases, which corresponded to the following equations: $s_1(x) = 0.2\sin(11.28x) + 1.02$ and $s_2(x) = 0.2\sin(11.28x + \pi) + 0.63$. These functions were the only ones for which the extrapolated points at x = 1.3 and 1.6 were not monotonically ordered (see Fig. 9). They were selected to examine if participants would correctly take into account the global oscillatory nature of the sine function.

For this experiment, the noise added for the generation of the actual scatterplots was kept fixed at 0.05 and the number of points was always 66. These values were chosen in order to allow participants to determine the nature of the curve without making it a trivial task. As in experiment 3, the location of the scatterplot was vertically jittered by a random amount (which was later corrected for in our analyses) and we included a considerable margin (12.5% of the screen) above and below the locations of the correct answers. Fig. 8 shows an example of stimulus for each generative function.

5.1.3. Procedure

Display parameters, procedure and task were the same as in experiment 3. Participants were explicitly asked to give an intuitive answer and to locate the point on their best estimate of the function from which the scatterplot was generated. The experimental design was a factorial design comprising all possible combinations of 5 generative functions, each with two possible signs of the derivative at x = 1, and two extrapolation positions, for a total of 20 conditions. The task was divided into 10 blocks, each comprising all 20 conditions in random order, for a total of 200 trials; the duration of each block was ~ 3 min. After each block, the participants could take a short break. No feedback was given. Before the beginning of the actual experiment, 25 practice trials were conducted under the supervision of the researcher.

5.2. Results

5.2.1. Location of the extrapolated point

First, we looked at the distribution of the participants' extrapolation responses on the y axis (Fig. 9). In all 20 conditions (except for quadratics, to which we later return), the center of the distribution fell close to the ideal answer (black dots in Fig. 9) and was clearly adapted to the function on the left. To verify this, we conducted a repeated measures ANOVA on participants' median extrapolation value with the type of function, the sign of the derivative and the probed position as within-participants factors, and we found a significant effect of function type ($F(2.65, 23.85) = 3.19$, partial $\eta^2 = 0.26$, $p < .05$), the sign of the derivative ($F(1,9) = 177.64$, Partial $\eta^2 = 0.95$, $p < .0001$), an interaction of the sign of the derivative with the probed position ($F(1,9) = 15.03$, partial $\eta^2 = 0.63$, $p < .01$), and a triple interaction of sign, probed position, and function type ($F(1.61, 14.52 = 33.39$, partial $\eta^2 = 0.79$, $p < .0001$). The effects and interactions involving function type indicated that participants varied their answers, not only according to the probed position (x = 1.3 or 1.6) or the derivative at the end point, but also, crucially, according to the type of function underlying the scatterplot. Since all functions ended with the same derivative at x = 1 (the rightmost point of the graph) this finding allows us to reject the hypothesis that participants were confined to a linear tangential extrapolation of the data.

Indeed, examination of the distributions made to the sinusoidal function made it clear that participants readily identified this function and gave adequate non-monotonic responses. In this condition only, the extrapolation at x = 1.6 significantly reverted and became closer to the graph mean (y = 0.825) relative to the extrapolation at x = 1.3 (for sinusoid with positive derivative: $\text{mean}_{x=1.3} = 1.20$, $\text{mean}_{x=1.6} = 0.87$, $t(9) = 6.89$, $p < 10^{-5}$; for sinusoid with negative derivative: $\text{mean}_{x=1.3} = 0.44$, $\text{mean}_{x=1.6} = 0.73$, $t(9) = -4.69$, $p = .001$). This observation is compatible with the oscillations of the sinusoidal function, but incompatible with any linear extrapolation, either based on a subset of the data points or on the tangent at x = 1. In all other conditions, the participants' extrapolations at x = 1.6 deviated more than those at x = 1.3, in agreement with the monotonicity of the underlying generative functions.

5.2.2. Response bias

We next examined how accurately the participants' extrapolations coincided with the ideal location, as derived from the functions used to generate the graphs. For linear and piecewise-linear functions, responses were biased towards extrapolations further away

from the expected point, as expected from Deming regression. 718 out of 1200 extrapolations resulted in regression lines steeper than the ideal response (59.83%, one sample proportion test: $\chi^2 = 46.413$, df = 1, $p < .0001$). Furthermore, examination of the mode of participants' responses revealed a systematic over-estimation of the absolute slope in 12 out of 12 combinations of sign, probed position, and function (see Fig. 9). Those results confirm the findings from experiment 3: participants' linear regressions are not unbiased, as would be expected under OLS assumptions, but are biased towards steeper slopes, as predicted by Deming-like regression.

For quadratic functions, participants' extrapolations were inaccurate, since their answers were considerably further from the expected location (see Fig. 9). One sample two-tailed t-tests on participants' median answers revealed a significant difference with the ideal answer for both convex and concave quadratic functions at the probed position of x = 1.6 ($mean_{convex} = 1.27$, $ideal_{convex} = 1.54$, t(9) = −4.39, p = .002 and $mean_{concave} = 0.47$, $ideal_{concave} = 0.11$, t(9) = 4.60, p = .001) and for the concave quadratic at the probed position of x = 1.3 (mean = 0.63, ideal = 0.55, t(9) = 2.656, p = .026); for the convex quadratic, the difference was in the proper direction but did not reach significance (mean = 1.07, ideal = 1.10, t(9) = −0.569, p = .6). Crucially, no significant difference was found between participants' answers and the answer expected if participants computed a linear extrapolation based on the derivative at x = 1 (dashed line in Fig. 9; all p values > 0.38).

5.2.3. Response variability

As we see from Fig. 9, the variability in responses greatly varied depending on the type of function and on the probed position; to investigate the statistical significance of this variation, we conducted a Fligner Killeen test, which is a test for homogeneity of variances most robust to departures from Gaussian distributions (see Conover, Johnson, & Johnson, 1981). It revealed no significant differences in the variance of the responses to functions having a positive or a negative derivative (FK med $X^2 = 0.137$, df = 1, p = 0.71). However, both the type of function and the probed x position had a significant impact (respectively: FK med $X^2 = 63.90$, df = 4, $p < 10^{-16}$ and FK med $X^2 = 94.37$, df = 9, $p < 10^{-16}$). Also, the variance of responses for piecewise linear functions with a late inflection was significantly higher than either the one for purely linear functions (FK med $X^2 = 14.33$, df = 1, $p < .001$) or for piecewise linear functions with an early inflection (FK med $X^2 = 6.63$, df = 1, p = .01). This increase in response variability is consistent with participants estimating a linear regression based on increasingly fewer data points (those on the right-hand side of the inflection for piecewise linear curves). It allows us to reject the hypothesis that participants used the fact that the piecewise linear functions were symmetrical, with the same absolute slope on both sides: if that was the case, there should have been no increase in variability, as the same number of points would have been available to estimate the slope for both types of piecewise linear functions. Crucially, no difference in responses' variance was found between quadratic and piecewise linear functions with a late inflection (FK med $X^2 = 0.004$, df = 1, p = .95). This finding, together with participants' inaccuracy for quadratics, suggests that quadratics might have been misperceived as ending with a linear trend.

To further confirm these findings, we conducted an ANOVA on the standard deviation of the participants' responses, with the type of function, the sign of the derivative and the probed position as within-participants factors. There was no effect of the sign of the derivative ($F(1, 9) = 0.02$, partial $\eta^2 = 0.003$, p = .88) or its interactions, indicating that participants treated symmetrically the upward and downward-going functions. As expected, we observed a main effect of the probed position ($F(1, 9) = 35.78$, partial $\eta^2 = 0.80$, $p < .001$), indicating that the standard deviation increased when the probed position went from x = 1.3 to x = 1.6, i.e. with greater extrapolation distance ($mean_{x=1.3} = 0.159$, $mean_{x=1.6} = 0.276$). Furthermore, there was a significant effect of the type of function ($F(2.63, 23.70) = 10.10$, partial $\eta^2 = 0.53$, $p < .001$) and its interaction with probed position ($F(2.48, 22.31) = 3.57$, partial $\eta^2 = 0.28$, p = .04). Post-hoc Tukey tests on the standard deviation of participants' given responses confirmed that response variability increased from linear (mean = 0.08) to both early (mean = 0.18, $p < .001$) and late (mean = 0.20, $p < .0001$) piecewise linear functions. It also significantly increased from linear to quadratic (mean = 0.18, $p < .01$) but not to sinusoid functions (mean = 0.12, p = .41). Also, we found no difference between the standard deviation of participants' answers for quadratic functions and for early (p = .94) and late (p = .78) piecewise linear functions.

5.2.4. Modelling of an optimal observer

One possible explanation of the poor performance with quadratics is that, given the noise level, there might not have been sufficient evidence to distinguish them from piecewise-linear functions. To clarify this point, we modeled a Bayesian optimal observer capable of selecting the best-fitting function within four families of functions: linear, piecewise-linear, quadratic or sinusoid. To find the best-fitting curve, the algorithm first used a minimization algorithm to identify, separately for each family of functions, the parameters that minimize the sum of the squares of the vertical distances of each data point to the curve (classical least squares). As a forward model, the algorithm assumed (correctly) that distance to the curve was drawn at random from a Gaussian distribution with standard deviation σ. Thus, the total Log likelihood of a graph was

$$ln(L) = -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n} {d_i}^2$$

where d_i are the vertical distances of each data point to the curve (i.e. $d_i = y_i - y_{model}$).

For the best-fitting model within each family of functions, we computed $\ln(L)$ using the above formula, with σ equal to the standard deviation of the residual distances. Finally, we selected the function for which the Bayesian Information Criterion (BIC) was minimal, i.e. the one which achieved an optimal trade-off between the number of free parameters (k) and the fit of the data points. We used the following formula: $BIC = k\ln(n) - 2ln(L)$, where k is the number of estimated parameters in a given model (including the fixed variance term, i.e. k = 3 for linear regression), and ln(L) is the likelihood of the data for that model. We preferred BIC over AIC since it has been

shown that BIC is asymptotically consistent, which means that it will select the true model if it figures among the models considered (Vrieze, 2012); also, BIC tends to penalize complex models more than AIC (Kuha, 2004).

We applied the model to the stimuli presented to the participants and we found that it was almost always able to select the correct underlying generative function (overall accuracy = 93.5%), even for datasets generated from quadratic functions (accuracy = 94.25%). This is crucial, since it means that the noise in our graphs was sufficiently low to make such a task possible. The optimal observer reproduced some, but not all of the features of the data. Like human participants, its response variability increased with the extrapolation distance, from x = 1.3 to x = 1.6. Like humans, the ideal observer was more precise for fully linear than for early-piecewise-linear and late-piecewise-linear functions. However, unlike humans, its estimates were unbiased and centered on the ideal location (see supplementary figure S2A). To reproduce human biases, we returned to Deming-like regression by replacing d_i in the above equations with the Euclidean distance of each data point to the nearest point of the curve. This procedure reduces to Deming regression when the function is linear, but extends the concept to any arbitrary function. When the sum of the squares of this Euclidean distance was minimized, the model, like human participants, revealed a bias towards extrapolations further away from the ideal point, in agreement with Deming regression (see supplementary figure S2B).

The last important misfit of the model was that it failed to reproduce the observed human inaccuracy with quadratic functions. Indeed, as expected from Deming regression, the mean extrapolations of the model for the quadratic function fell slightly *beyond* the ideal ones (e.g. for the convex quadratic and x = 1.6, correct extrapolation = 1.54, mean model response = 1.57), whereas the converse was true for our participants (mean response = 1.26, which is quite close to the linear tangent-based extrapolation = 1.22). We therefore considered a model that did not include the quadratic functions as one of the possible fits. In this case, quadratic functions were classified as piecewise linear functions 100% of the time. However, as seen in supplementary figure S2C, the fit to human data remained inadequate, since the model now predicted values that were almost constant as a function of x position, and way too close to the mean of the data (e.g. for the convex quadratic and x = 1.6, mean response = 0.96).

Note, however, that for quadratics, the participants' responses exhibited a large variance and a distribution with multiple peaks roughly coinciding with the three above possibilities (constant response, linear tangent, or quadratic extrapolation); it is therefore possible that they adopted a mixture of the above strategies, and/or that they correctly identified the quadratic but failed to correctly follow its curvature and, instead, based their extrapolations on the tangent line. We can at least conclude from our theoretical analysis that (1) all functions were clearly discriminable, including the quadratic and piecewise-linear functions; and yet (2) participants performed poorly only with quadratic functions.

5.3. Discussion

Experiment 4 showed that, unlike a fixed linear regression package, participants could flexibly adapt their regressions to the nature of the graphs they were exposed to. They understood linearity, bi-linearity, and periodicity (sine function) and used this knowledge to extrapolate points outside the observational range, even for non-linear functions. The sole exception was quadratic functions, for which participants systematically underestimated the size of the variations of the function on most trials, and often chose to follow the tangential line rather than the actual curvature of the quadratic. This finding is consistent with prior evidence for a human preference for linear trends (Kalish et al., 2007; Little & Shiffrin, 2009; Mcdaniel & Busemeyer, 2005). Several previous studies reported a so-called "exponential growth bias", according to which humans, when facing a series of data points that underlie an exponential increase, tend to underestimate the position of a point outside the observational range (Andreassen & Kraus, 1990; Eggleton, 1982; Wagenaar & Timmers, 1978; Wagenaar & Sagaria, 1975; Wagenaar & Timmers, 1979). Our results uncovered such an underestimation for quadratic functions as well, thus suggesting that it might occur more generally for curved functions, not just exponentials. Whether this bias is proportional to the function's acceleration should be investigated by future studies. Also, we show that this bias may be restricted to non-linear functions that are, in their extrapolation area, monotonic, since participants correctly extrapolated sinusoids with a high degree of precision.

6. General discussion

We start by summarizing the results. Across four experiments probing the perception of noisy scatterplots, educated human adults exhibited a remarkable ability for "mental regression". Regardless of the task (trend judgment, line fitting or extrapolation), the presentation time of the stimuli (100 ms or long exposure) and the nature of the generative function (linear or non-linear), their performance closely tracked that of an optimal statistical model. Their performance was tightly correlated to the *t* test associated to the Pearson coefficient of correlation of the dataset, and thus took into account various aspects of the available statistical evidence, including the noise level and the number of data points. Surprisingly, however, their mental regression procedure exhibited a strong bias towards exaggerating the regression slope, thus departing from Ordinary Least Squares (OLS) regression and coming closer to Deming regression (where what is minimized is the Euclidean distance of the data points to the regression curve). The Deming-like bias was replicated multiple times in both line-estimation and extrapolation tasks. We also found that human extrapolation abilities were not restricted to linear relationships, but extended to scatterplots underlying piecewise-linear functions (with both an early and a late inflection), sinusoids, but not quadratics, whose curvature seems to be consistently underestimated. In this discussion, we examine several points raised by these findings: possible explanations for the Deming bias; its impact on real-life graphical analyses; the misperception of quadratic functions; the integration of our findings within the field of function learning; and finally, perspectives and suggestions for future studies in the field of graphicacy.

6.1. The Deming bias

How could we explain that humans use Deming regression rather than ordinary least squares, although the latter procedure would have been more adequate for our datasets, where the y values were noisy measurements of fixed x values? A Deming-like regression may be rational for several reasons. First, Deming is the appropriate procedure when the x measurements are noisy – and participants may not have been aware that x values were fixed. Second, Deming emerges naturally if participants treat the x and y axes symmetrically, which results in minimizing the distance of the points to the fit from both points' coordinates. Indeed, the asymmetry in OLS regression (resulting in a different regression line when y is regressed on x or vice-versa) is highly counter-intuitive. Our results show that humans spontaneously perceive the principal axis of the graph, which is defined as the straight line that minimizes the sum of the squared Euclidean distances to the set of points, and hence corresponds exactly to the Deming regression line. Interestingly, considerable research indicates that, during both the perception of objects and of the geometry of the environment, humans and other animals spontaneously extract the principal axis of simple shapes and use it in various computations including object perception (Cohen & Singh, 2006), object misperception in patients with hemineglect (Driver, Baylis, Goodrich, & Rafal, 1994), object manipulation (Turvey, Burton, Pagano, Solomon, & Runeson, 1992), grasping (Cuijpers, Smeets, & Brenner, 2004), visual search (Boutsen & Marendaz, 2001), or spatial reorientation (Bodily et al., 2011, 2018; Cheng, 2005). It therefore makes sense that, when confronted with the task of perceiving the direction of a scatterplot, human adults would spontaneously reuse this evolutionarily ancient ability to grasp an object's principal axis. According to this hypothesis, graph perception would constitute a novel instance of "neuronal recycling" (Dehaene, 2005; Dehaene & Cohen, 2007), i.e. the repurposing of pre-existing and evolutionary older cognitive mechanisms, initially devoted to other purposes, for novel cultural uses. Just like the invention of writing repurposes part of the ventral visual system for object recognition towards the fast recognition of letter strings (Dehaene, 2009), the cultural invention of the scatter plot could be viewed as a clever way, starting from a large set of numerical data points, to display them in 2D or 3D space such that the resulting graph benefits from the human visual system's sophisticated parallel processing ability, resulting in an immediate extraction of its principal axis.

Interestingly, object perception is not limited to the perception of a single straight axis (the principal axis), but can accommodate the perception of complex objects or spaces comprising multiple, possibly flexible subparts. In the case, the visual system appears to extract multiple axes of quasi-symmetry that may correspond to the mathematical concept of "medial axis" or "shape skeleton" (Ayzenberg & Lourenco, 2019; Cohen & Singh, 2006; Kelly & Durocher, 2011; Kovacs & Julesz, 1994; Lowet, Firestone, & Scholl, 2018). Mathematically, a medial axis is the locus of points equidistant from the two closest shape boundaries (Blum, 1967, 1973). The concept of medial axis generalizes the notion of principal axis to curved shapes (e.g. the skeleton of a snake). As such, it may explain why, in experiment 4, we found that human participants could extrapolate the scatter plots of non-linear functions, an ability that goes beyond the mere extraction of the principal axis. The capacity to approximate the medial axis of a graph, i.e. the locus of the curve that minimizes the sum of the square Euclidean distances to the data points, would readily explain how humans extract non-linear curves from scatter plots, and why they show a Deming-like bias when doing so. More studies will be needed to further test the reliance of participants on medial axis extraction and, more broadly, to verify to what extent the neural recycling proposal applies to graphicacy.

6.2. Impact of Deming bias in real-life graphical analyses

Regardless of its ultimate cause, the fact that human adults compute a Deming rather than an OLS regression may have considerable implications in real-life uses of graphical representations, such as in finance, where stock markets' noisy graphs are often used by investors to make quick selling or buying decisions. Biases in economic and financial behavior have typically been reduced to various cognitive biases (Kahneman, 2003; Ricciardi & Simon, 2000), such as confirmation bias (Nickerson, 1998) or loss aversion (Kahneman, Knetsch, & Thaler, 1991). Our findings suggest that such biases, although certainly at play, might not be the only factors influencing financial behavior. The Deming bias implies that investors could be more likely to keep investing in stocks showing an uprising trend (or selling stocks revealing a negative trend), because they perceive the trend as steeper than it actually is. Indeed, finance experts strongly rely on the slope information when looking at a graph (Beattie & Jones, 2002), and it is known that data series presented in graphical rather than tabular forms generally lead to a worse understanding of the actual trend in a dataset (DeLosh et al., 1997; Lawrence & Makridakis, 1989). Further studies of graph perception could be performed with specific populations such as finance experts (using both tabular and graphical forms) in order to disentangle the biases of geometrical origin from other reasoning and cognitive biases, and to measure the practical import of the present findings.

6.3. The misperception of quadratic functions.

Aside from the Deming bias, our series of studies demonstrates the high sophistication of human graph perception. Previous research by Schultz and collaborators (2017) elegantly showed that human adults can infer the complex functions underlying a mathematical graph, as long as those functions are compositional, i.e. can be decomposed into simpler constituents such as linear or oscillatory trends. Our results extend this ability to noisy scatterplots, which prevent participants from directly perceiving the function and force them to extract it from noise. While participants were generally accurate, experiment 4 showed that quadratics were consistently misperceived. This finding is compatible with the hypothesis that human compositional grammar (Piantadosi, Tenenbaum, & Goodman, 2016) may be limited to a certain set of primitives such as linear and oscillating functions (Schulz et al., 2017) but not quadratics. It fits with previous observations that exponentially accelerating progressions are consistently underestimated (Andreassen & Kraus, 1990; Eggleton, 1982; Wagenaar & Timmers, 1978; Wagenaar & Sagaria, 1975; Wagenaar & Timmers, 1979).

Tentatively, the misperception of accelerated functions may also be related to the history of science and, specifically, the remarkably slow discovery of the quadratic law of falling bodies. For centuries, the accepted theory of the "impetus" misconstrued the trajectory of a projectile as a line, followed by an arc of a circle, followed by a vertical fall. Galileo finally managed to correctly describe the projectile motion as a quadratic function and formalized the concept of uniform acceleration of a falling body, which is crucial in Newton's second law of dynamics. Interestingly, it has been shown that most students hold pre-Newtonian intuitive views of motion (Espinoza, 2005; McCloskey, Washburn, & Felch, 1983). The misperception of quadratics might therefore be considered as part of a more general misunderstanding of the concepts of acceleration and deceleration. It would be interesting to further investigate the acquisition of quadratic functions in mathematics students and the specific difficulties that it raises.

6.4. Integration of our findings within the literature on function learning

In our series of studies, we used scatterplots as stimuli, in order to specifically investigate human ability to extract statistical information from simultaneously presented bivariate data. Therefore, our findings primarily contribute to the field of graph perception. Our results, however, might also be integrated within the larger body of literature on function learning (Kalish et al., 2004; Lewandowsky et al., 2002; Lucas et al., 2015; Mcdaniel & Busemeyer, 2005). As mentioned in the introduction, the experimental procedure and stimuli are quite different. Indeed, our experimental procedure offers a complementary methodology to function learning studies that mostly relied on long training sessions involving the implicit learning of functional associations between serially presented data pairs. Our studies demonstrate the great advantage of scatterplots, which allow for the rapid extraction of a mathematical relationship hidden in a noisy graphical dataset, in less than one second, without need to serially map and update the association between input and output values (Bott & Heit, 2004; Carroll, 1963; DeLosh et al., 1997; Kalish, 2013; Kalish et al., 2004), thus avoiding any recourse to memory processes. In spite of such differences, however, our findings converge in confirming a human predisposition towards linear functions (Kalish et al., 2004), which were processed more accurately than other bilinear, sinusoidal, or especially quadratic functions. The distribution of responses we observed in the extrapolation of quadratic functions (exemplified by the three "bumps" in Fig. 9, fourth row) might be seen as supporting the idea that humans possess multiple concurrent extrapolation methods (one being a linear extrapolation), as proposed in recent models of function learning (Kalish et al., 2004; Lewandowsky et al., 2002). Our results also show that complex non-linear relationships, such as sinusoidal functions, which were found particularly hard to encode in classic function learning studies (DeLosh et al., 1997; Kalish, 2013), are nevertheless available to human adults when the data is presented as a scatterplot. This point confirms and extends previous literature that used similar graphical stimuli (Little & Shiffrin, 2009; Schulz et al., 2015, 2017) and suggests that the perception of Cartesian plots may be a method of choice in order to explore the limits of human function learning.

6.5. Open questions

Our studies leave open many issues for further investigations. First, how robust is human mental regression? Can humans reject outliers in a scatterplot? What is the range of mathematical functions that humans can readily perceive in a graph? Graphicacy should also be investigated in its developmental and neural aspects: how do students acquire graph perception abilities, and to what extent can their competence be improved by training? And where in the brain are scatterplots processed? Does their perception recruit the vast network areas involved in calculation and higher mathematics or, as the neuronal recycling would suggest, does it mobilize a restricted subset of areas for shape perception? In the future, those issues could be studied in both educated and naïve adults and children.

Acknowledgements

Supported by INSERM, CEA, Collège de France, Foundation Bettencourt-Schueller, French Ministry of Education, Ecole Doctorale FIRE, an ERC grant "NeuroSyntax" to S.D and a Mind Science Foundation grant to L.C. We are particularly grateful to Marie Lubineau, Mathias Sablé-Meyer and Guillaume Dehaene for extensive discussions of experimental design and data analysis.

Appendix A. Optimal decision making in the trend judgment task

According to classical statistical theory (David & Neyman, 1938; Theil, 1971), given a dataset generated according to a noisy linear function (i.e. n pairs of data points $\{x_i, y_i\}$, where $y_i = \alpha x_i + \varepsilon_i$ and the ε_i are independent centered Gaussian samples with standard deviation σ), the best linear unbiased estimator (BLUE) of the slope α is

$$\widehat{\alpha} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

According to the Gauss-Markov theorem, this estimator is unbiased (its mean is equal to α) and has minimum variance when compared to other possible estimators. Thus, it is the most appropriate estimator on which to base the trend judgment task. In order to optimally decide whether the trend in the graph is increasing or decreasing, an ideal observer should base its decision on whether the slope estimate $\widehat{\alpha}$ is positive or negative.

Assuming now that this is the decision strategy, can we predict how the associated error rate and response times should vary with experimental parameters? According to the tenets of classical signal detection theory (Green & Swets, 1966) and its extension to response times (e.g., Gold & Shadlen, 2002) the difficulty and error rate of such a decision is determined not only by the mean of this

variable, but also by its distribution across trials. The standard error of the slope estimate $\widehat{\alpha}$ is given by

$$s_{\widehat{\alpha}} = \sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Under the null hypothesis, the ratio of the slope estimate to its standard error has a Student's t-distribution with $n - 2$ degrees of freedom (i.e. a distribution close to a Gaussian for large n). This t value can also be written as

$$t = \frac{\widehat{\alpha}}{s_{\widehat{\alpha}}} = \sqrt{n-2}\ \frac{r}{\sqrt{1-r^2}}$$

where r is the Pearson coefficient of correlation, given by the covariance of the x and y values divided by the product of their standard deviations (for a comprehensive explanation, see Baguley, 2012):

$$r = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

The t value is the basis for the classical statistical test for significance of a linear trend: to decide if a non-null (positive or negative) trend is present, we compare the observed t value to the Student's t distribution expected under the null hypothesis. Here, however, the situation is a bit different: as an experimenter, we know how the data was generated on each trial with a given slope α, which may be different from zero; and, to compute the psychometric function, we would like to know what is the probability that the decision maker will respond "the trend is increasing", assuming that the decision is based on whether $\widehat{\alpha}$ is greater than zero. Under these conditions, the t value is no longer distributed as a Student's t-distribution (because the expected value of $\widehat{\alpha}$, being an unbiased estimator, is α). However, the following value, $t' = \frac{\widehat{\alpha}-\alpha}{s_{\widehat{\alpha}}}$, is again distributed as a Student's t-distribution. Thus, the probability of the "increasing" response is:

$$p_{increasing} = p(\widehat{\alpha} > 0) = p\left(t' s_{\widehat{\alpha}} + \alpha > 0\right) = p\left(t' > \frac{-\alpha}{s_{\widehat{\alpha}}}\right) = p(t' > -t) = \int_{-t}^{+\infty} Student_n(u)du = \int_{-\infty}^{t} Student_n(u)du$$

This equation indicates that the proportion of responses is an increasing function of the t value. More specifically, it is a sigmoid-like function, namely the cumulative Student's t-distribution. Note that, strictly speaking, this function still depends on the number of points n. However, as n increases, it quickly becomes essentially indistinguishable from the integral of a Gaussian, and hence independent of n; it is also extremely similar to the classical sigmoid (see Gold & Shadlen, 2002).

The above theory, analogous to classical signal detection theory (SDT; Green & Swets, 1966), assumes that the decision is based on a single sample of t, and predicts only the psychometric function (or, equivalently, error rates) but not response times. To predict response times, we turn to a "sequential probability ratio test" variant of the above theory, according to which the decision-maker accumulates noisy samples of evidence about the sign of $\widehat{\alpha}$, up to a fixed decision bound. Under such an accumulation-of-evidence model, according to the equations in (Gold & Shadlen, 2002), the psychometric response function becomes the classic sigmoid:

$$p_{increasing} = \frac{1}{1 + e^{-2Bt}}$$

And the response time is predicted by the deviation of the absolute value of t from zero, according to a decreasing, convex upward function given by the equation:

$$RT = \frac{B}{|t|}\tanh(B|t|)$$

In those equations, B is a constant that jointly reflects both the sensitivity of the decision-maker (the amount of information accumulated per unit of time) and his decision threshold (controlling the speed/accuracy tradeoff).

In summary, the theory predicts that decision difficulty (both error rate and RT) should be determined by the t value. The equation for t, in turn, makes it clear that the decision difficulty should depend on all three of the manipulated graph parameters (n, σ and α), and predicts that the effects of these variables should be jointly summarized by a single effect of the t value on behavior.

Appendix B. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cogpsych.2021.101406.

References

Andreassen, P. B., & Kraus, S. J. (1990). Judgmental extrapolation and the salience of change. *Journal of Forecasting, 9*(4), 347–372. https://doi.org/10.1002/for.3980090405.

Appelle, S. (1972). Perception and discrimination as a function of stimulus orientation: The "oblique effect" in man and animals. *Psychological Bulletin, 78*(4), 266.

Ayzenberg, V., & Lourenco, S. F. (2019). Skeletal descriptions of shape provide unique perceptual information for object recognition. *Scientific Reports, 9*(1), 9359. https://doi.org/10.1038/s41598-019-45268-y.
Baguley, T. (2012). *Serious stats: A guide to advanced statistics for the behavioral sciences*. Macmillan International Higher.
Balchin, W. G. V., & Coleman, A. M. (1966). Graphicacy should be the fourth ace in the pack. *Cartographica: The International Journal for Geographic Information and Geovisualization, 3*(1), 23–28. https://doi.org/10.3138/C7Q0-MM01-6161-7315.
Beattie, V., & Jones, M. J. (2002). The Impact of Graph Slope on Rate of Change Judgments in Corporate Reports. *Abacus, 38*(2), 177–199. https://doi.org/10.1111/1467-6281.00104.
Blum, H. (1967). *A transformation for extracting new descriptors of shape* (p. 4). Cambridge: MIT Press.
Blum, H. (1973). Biological shape and visual science (part I). *Journal of Theoretical Biology, 38*(2), 205–287. https://doi.org/10.1016/0022-5193(73)90175-6.
Bobko, P., & Karren, R. (1979). The perception of Pearson product moment correlations from bivariate scatterplots. *Personnel Psychology, 32*(2), 313–325. https://doi.org/10.1111/j.1744-6570.1979.tb02137.x.
Bodily, K. D., Eastman, C. K., & Sturz, B. R. (2011). Neither by global nor local cues alone: Evidence for a unified orientation process. *Animal Cognition, 14*(5), 665–674. https://doi.org/10.1007/s10071-011-0401-x.
Bodily, K. D., Sullens, D. G., Price, S. J., & Sturz, B. R. (2018). Testing principal- versus medial-axis accounts of global spatial reorientation. *Journal of Experimental Psychology: Animal Learning and Cognition, 44*(2), 209–215. https://doi.org/10.1037/xan0000162.
Bolger, F., & Harvey, N. (1993). Context-Sensitive Heuristics in Statistical Reasoning. *The Quarterly Journal of Experimental Psychology Section A, 46*(4), 779–811. https://doi.org/10.1080/14640749308401039.
Bott, L., & Heit, E. (2004). Nonmonotonic Extrapolation in Function Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(1), 38–50. https://doi.org/10.1037/0278-7393.30.1.38.
Boutsen, L., & Marendaz, C. (2001). Detection of shape orientation depends on salient axes of symmetry and elongation: Evidence from visual search. *Perception & Psychophysics, 63*(3), 404–422. https://doi.org/10.3758/BF03194408.
Carroll, J. D. (1963). Functional learning: The learning of continuous functional mappings relating stimulus and response continua. *ETS Research Bulletin Series, 1963*(2), i–144. https://doi.org/10.1002/j.2333-8504.1963.tb00958.x.
Cheng, K. (2005). Reflections on geometry and navigation. *Connection Science, 17*(1–2), 5–21. https://doi.org/10.1080/09540090500138077.
Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research, 43*(4), 393–404.
Chong, S. C., & Treisman, A. (2005). Statistical processing: Computing the average size in perceptual groups. *Vision Research, 45*(7), 891–900.
Ciccione, L., & Dehaene, S. (2020). Grouping Mechanisms in Numerosity Perception. *Open Mind, 4*, 102–118. https://doi.org/10.1162/opmi_a_00037.
Cleveland, W., Diaconis, P., & Mcgill, R. (1982). Variables on Scatterplots Look More Highly Correlated When the Scales Are Increased. *Science, 216*(4550), 1138–1141. https://doi.org/10.1126/science.216.4550.1138.
Cleveland, W., & McGill, R. (1984). Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association, 79*(387), 531–554. https://doi.org/10.1080/01621459.1984.10478080.
Cleveland, W., & McGill, R. (1985). Graphical Perception and Graphical Methods for Analyzing Scientific Data. *Science, New Series, 229*(4716), 828–833.
Cleveland, W., & McGill, R. (1987). Graphical Perception: The Visual Decoding of Quantitative Information on Graphical Displays of Data. *Journal of the Royal Statistical Society. Series A (General), 150*(3), 192. https://doi.org/10.2307/2981473.
Cohen, E. H., & Singh, M. (2006). Perceived orientation of complex shape reflects graded part decomposition. *Journal of Vision, 6*(8). https://doi.org/10.1167/6.8.4.
Conover, W. J., Johnson, M. E., & Johnson, M. M. (1981). A Comparative Study of Tests for Homogeneity of Variances, with Applications to the Outer Continental Shelf Bidding Data. *Technometrics, 23*(4), 351–361. https://doi.org/10.1080/00401706.1981.10487680.
Correll, M., & Heer, J. (2017). Regression by Eye: Estimating Trends in Bivariate Visualizations. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17, 1387–1396. Doi: 10.1145/3025453.3025922.
Cuijpers, R. H., Smeets, J. B. J., & Brenner, E. (2004). On the Relation Between Object Shape and Grasping Kinematics. *Journal of Neurophysiology, 91*(6), 2598–2606. https://doi.org/10.1152/jn.00644.2003.
Curcio, F. R. (1987). Comprehension of Mathematical Relationships Expressed in Graphs. *Journal for Research in Mathematics Education, 18*(5), 382. https://doi.org/10.2307/749086.
David, F. N., & Neyman, J. (1938). Extension of the Markoff Theorem on Least Squares. *Statistical Research Memoirs, 2*, 105–116.
Dehaene, S. (2009). *Reading in the brain*. Penguin Viking.
Dehaene, S. (2005). Evolution of human cortical circuits for reading and arithmetic: The "neuronal recycling" hypothesis. In From monkey brain to human brain (p. 33).
Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General, 122*(3), 371–396. https://doi.org/10.1037/0096-3445.122.3.371.
Dehaene, S., & Cohen, L. (2007). Cultural Recycling of Cortical Maps. *Neuron, 56*(2), 384–398. https://doi.org/10.1016/j.neuron.2007.10.004.
DeLosh, E. L., McDaniel, M. A., & Busemeyer, J. R. (1997). Extrapolation: The Sine Qua Non for Abstraction in Function Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(4), 968–986.
Deming, W. E. (1943). *Statistical adjustment of data*. Wiley.
Driver, J., Baylis, G. C., Goodrich, S. J., & Rafal, R. D. (1994). Axis-based neglect of visual shapes. *Neuropsychologia, 32*(11), 1353–1356. https://doi.org/10.1016/0028-3932(94)00068-9.
Eggleton, I. R. C. (1982). Intuitive Time-Series Extrapolation. *Journal of Accounting Research, 20*(1), 68. https://doi.org/10.2307/2490763.
Espinoza, F. (2005). An analysis of the historical development of ideas about motion and its implications for teaching. *Physics Education, 40*(2), 139–146. https://doi.org/10.1088/0031-9120/40/2/002.
Friel, S. N., Curcio, F. R., & Bright, G. W. (2001). Making Sense of Graphs: Critical Factors Influencing Comprehension and Instructional Implications. *Journal for Research in Mathematics Education, 32*(2), 124. https://doi.org/10.2307/749671.
Gold, J. I., & Shadlen, M. N. (2002). Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward. *Neuron, 36*(2), 299–308.
Green, D., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Krieger Publishing Company.
Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24*(2), 95–112. https://doi.org/10.1007/BF02289823.
Harrison, L., Yang, F., Franconeri, S., & Chang, R. (2014). Ranking Visualizations of Correlation Using Weber's Law. *IEEE Transactions on Visualization and Computer Graphics, 20*(12), 1943–1952. https://doi.org/10.1109/TVCG.2014.2346979.
Harvey, N., Ewart, T., & West, R. (1997). Effects of Data Noise on Statistical Judgement. *Thinking & Reasoning, 3*(2), 111–132. https://doi.org/10.1080/135467897394383.
Kahneman, D. (2003). Maps of Bounded Rationality: Psychology for Behavioral Economics. *American Economic Review, 93*(5), 1449–1475. https://doi.org/10.1257/000282803322655392.
Kahneman, D., Knetsch, J., & Thaler, R. (1991). Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias. *Journal of Economic Perspectives, 5*(1), 193–206.
Kalish, M. L. (2013). Learning and extrapolating a periodic function. *Memory & Cognition, 41*(6), 886–896. https://doi.org/10.3758/s13421-013-0306-9.
Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review, 14*(2), 288–294. https://doi.org/10.3758/BF03194066.
Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of Linear Experts: Knowledge Partitioning and Function Learning. *Psychological Review, 111*(4), 1072–1099. https://doi.org/10.1037/0033-295X.111.4.1072.
Kelly, D. M., & Durocher, S. (2011). Comparing geometric models for orientation. *Communicative & Integrative Biology, 4*(6), 710–712.
Kosslyn, S. M., & Kosslyn, S. M. (2006). *Graph Design for the Eye and Mind*. USA: Oxford University Press.
Kovacs, I., & Julesz, B. (1994). Perceptual sensitivity maps within globally defined visual shapes. *Nature, 370*(6491), 644–646. https://doi.org/10.1038/370644a0.

Kuha, J. (2004). AIC and BIC: Comparisons of Assumptions and Performance. *Sociological Methods & Research, 33*(2), 188–229. https://doi.org/10.1177/0049124103262065.
Lane, D. M., Anderson, C. A., & Kellam, K. L. (1985). Judging the Relatedness of Variables: The Psychophysics of Covariation Detection. *Journal of Experimental Psychology: Human Perception and Performance, 11*(5), 640.
Lawrence, M., Goodwin, P., O'Connor, M., & Önkal, D. (2006). Judgmental forecasting: A review of progress over the last 25years. *International Journal of Forecasting, 22*(3), 493–518. https://doi.org/10.1016/j.ijforecast.2006.03.007.
Lawrence, M., & Makridakis, S. (1989). Factors affecting judgmental forecasts and confidence intervals. *Organizational Behavior and Human Decision Processes, 43*(2), 172–187. https://doi.org/10.1016/0749-5978(89)90049-6.
Lewandowsky, S., Kalish, M., & Ngang, S. K. (2002). Simplified learning in complex situations: Knowledge partitioning in function learning. *Journal of Experimental Psychology: General, 131*(2), 163–193. https://doi.org/10.1037/0096-3445.131.2.163.
Linnet, K. (1998). Performance of Deming regression analysis in case of misspecified analytical error ratio in method comparison studies. *Clinical Chemistry, 44*(5), 1024–1031. https://doi.org/10.1093/clinchem/44.5.1024.
Little, D. R., & Shiffrin, R. M. (2009). Simplicity Bias in the Estimation of Causal Functions. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (p. 7).
Lowet, A. S., Firestone, C., & Scholl, B. J. (2018). Seeing structure: Shape skeletons modulate perceived similarity. *Attention, Perception, & Psychophysics, 80*(5), 1278–1289. https://doi.org/10.3758/s13414-017-1457-8.
Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic Bulletin & Review, 22*(5), 1193–1215. https://doi.org/10.3758/s13423-015-0808-5.
Martin, R. F. (2000). General Deming Regression for Estimating Systematic Bias and Its Confidence Interval in Method-Comparison Studies. *Clinical Chemistry, 46*(1), 100–104. https://doi.org/10.1093/clinchem/46.1.100.
McCloskey, M., Washburn, A., & Felch, L. (1983). Intuitive physics: The straight-down belief and its origin. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9*(4), 636–649. https://doi.org/10.1037/0278-7393.9.4.636.
Mcdaniel, M. A., & Busemeyer, J. R. (2005). The conceptual basis of function learning and extrapolation: Comparison of rule-based and associative-based models. *Psychonomic Bulletin & Review, 12*(1), 24–42.
McElroy, F. W. (1967). A Necessary and Sufficient Condition That Ordinary Least-Squares Estimators Be Best Linear Unbiased. *Journal of the American Statistical Association, 62*(320), 1302–1304. https://doi.org/10.1080/01621459.1967.10500935.
Mosteller, F., Siegel, A. F., Trapido, E., & Youtz, C. (1981). Eye-Fitting of Straight Lines. *The American Statistician, 2*, 150–152.
Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology, 2*(2), 175–220.
Parrott, S., Guzman-Martinez, E., Ortega, L., Grabowecky, M., Huntington, M. D., & Suzuki, S. (2014). Spatial Position Influences Perception of Slope from Graphs. *Perception, 43*(7), 647–653. https://doi.org/10.1068/p7758.
Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2016). The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological Review, 123*(4), 392–424. https://doi.org/10.1037/a0039980.
Poulton, E. C. (1985). Geometric illusions in reading graphs. *Perception & Psychophysics, 37*(6), 543–548. https://doi.org/10.3758/BF03204920.
Puntanen, S., & Styan, G. P. H. (1989). The Equality of the Ordinary Least Squares Estimator and the Best Linear Unbiased Estimator. *The American Statistician, 43*(3), 153–161. https://doi.org/10.1080/00031305.1989.10475644.
Rensink, R. A., & Baldridge, G. (2010). The Perception of Correlation in Scatterplots. *Computer Graphics Forum, 29*(3), 1203–1210. https://doi.org/10.1111/j.1467-8659.2009.01694.x.
Ricciardi, V., & Simon, H. (2000). What is behavioral finance? *Business, Education & Technology Journal, 2*(2), 1–9.
Schulz, E., Tenenbaum, J. B., Duvenaud, D., Speekenbrink, M., & Gershman, S. J. (2017). Compositional inductive biases in function learning. *Cognitive Psychology, 99*, 44–79. https://doi.org/10.1016/j.cogpsych.2017.11.002.
Schulz, E., Tenenbaum, J., Reshef, D., Speekenbrink, M., & Gershman, S. (2015). Assessing the Perceived Predictability of Functions. *CogSci, 6*.
Spence, I. (2006). William Playfair and the Psychology of Graphs. Proceedings of the American Statistical Association, Section on Statistical Graphics, 2426–2436.
Strahan, R. F., & Hansen, C. J. (1978). Underestimating Correlation from Scatterplots. *Applied Psychological Measurement, 2*(4), 543–550. https://doi.org/10.1177/014662167800200409.
Surber, C. (1986). Model Testing Is Not Simple: Comments on Lane, Anderson, and Kellam. *Journal of Experimental Psychology: Human Perception and Performance, 12*(1), 108–109.
Theil, H. (1971). *Principles of Econometrics*. New York: John Wiley and Sons.
Tufte, E. (2001). *The visual display of quantitative information, Vol. 2*. Cheshire, CT: Graphics Press.
Turvey, M. T., Burton, G., Pagano, C. C., Solomon, H. Y., & Runeson, S. (1992). Role of the inertia tensor in perceiving object orientation by dynamic touch. *Journal of Experimental Psychology: Human Perception and Performance, 18*(3), 714–727. https://doi.org/10.1037/0096-1523.18.3.714.
Van Opstal, F., de Lange, F. P., & Dehaene, S. (2011). Rapid parallel semantic processing of numbers without awareness. *Cognition*. https://doi.org/10.1016/j.cognition.2011.03.005.
Vanderplas, S., Cook, D., & Hofmann, H. (2020). Testing Statistical Charts: What Makes a Good Graph? *Annual Review of Statistics and Its Application, 7*(1), 61–88. https://doi.org/10.1146/annurev-statistics-031219-041252.
Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods, 17*(2), 228–243. https://doi.org/10.1037/a0027127.
Wagenaar, W. A., & Timmers, H. (1978). Extrapolation of exponential time series is not enhanced by having more data points. *Perception & Psychophysics, 24*(2), 182–184. https://doi.org/10.3758/BF03199548.
Wagenaar, W., & Sagaria, S. (1975). Misperception of exponential growth. *Perception & Psychophysics, 18*(6), 416–422. https://doi.org/10.3758/BF03204114.
Wagenaar, W., & Timmers, H. (1979). The pond-and-duckweed problem; three experiments on the misperception of exponential growth. Acta Psychologica, 43, 239–251.