Journal of Experimental Psychology: Human Perception and Performance

Outlier Detection and Rejection in Scatterplots: Do Outliers Influence Intuitive Statistical Judgments?

Lorenzo Ciccione, Guillaume Dehaene, and Stanislas Dehaene

Online First Publication, November 17, 2022. https://dx.doi.org/10.1037/xhp0001065

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: Human Perception and Performance

ISSN: 0096-1523

https://doi.org/10.1037/xhp0001065

Outlier Detection and Rejection in Scatterplots: Do Outliers Influence Intuitive Statistical Judgments?

Lorenzo Ciccione[1, 2], Guillaume Dehaene[3], and Stanislas Dehaene[1, 2]

[1] UNICOG Cognitive Neuroimaging Lab, CEA, INSERM, Université Paris-Saclay, NeuroSpin Center

[2] Collège de France, Université Paris Sciences Lettres (PSL)

[3] Département des Neurosciences Fondamentales, University of Geneva

According to a growing body of research, human adults are remarkably accurate at extracting intuitive statistics from graphs, such as finding the best-fitting regression line through a scatterplot. Here, we ask whether humans can also perform outlier rejection, a nontrivial statistical problem. In three experiments, we investigated human adults' capacity to evaluate the linear trend of a flashed scatterplot comprising 0–4 outlier datapoints. Experiment 1 showed that participants did not spontaneously reject outliers: when outliers were not mentioned, their presence biased the participants' trend judgments and regression line estimates. In Experiment 2, where participants were explicitly asked to exclude outliers, the outlier-induced bias was reduced but remained significant. In Experiment 3, where participants were asked to explicitly detect any outlier before adjusting their regression line, outlier detection was satisfactory, but the detected outliers continued to bias the regression responses, unless they were quite distant from the main regression line. We propose a simple model for outlier detection, based on the computation of a *z*-score that estimates how far a given datapoint is from the distribution of distances to the regression line, and we show that this model closely approximates human performance. Detection is not rejection, however, and our results suggest that humans can remain biased by outliers that they have detected.

Public Significance Statement

In all fields of science, it is quite common that a handful of observations depart from the rest of a dataset. In statistical jargon, such exceptional observations are known as *outliers*. In most cases, they should be ignored in order to focus the analysis on the typical case—or at least they should be analyzed separately. In our study, we tested whether human adults can perceptually detect and discard outliers when attempting to intuitively extract the trend from a scatterplot. We find that, spontaneously, adults do not reject outliers and are therefore strongly influenced by them in their statistical judgements. Furthermore, even when adults are told to detect and reject outliers, they continue to be biased by them. We propose guidelines for graphics designers to facilitate outlier detection and rejection.

Keywords: graph perception, outliers, mental regression, attention, intuitive statistics

Supplemental materials: https://doi.org/10.1037/xhp0001065.supp

Lorenzo Ciccione https://orcid.org/0000-0002-5132-2531

We report our sample size determination, data exclusion, experimental manipulations and statistical analyses. All data and analysis' codes are available at https://osf.io/bfjsw/. Stimuli were generated using Python and data were analyzed using R. The study was not preregistered.

This work was supported by the Ministère de l'Education Nationale (France), the "FIRE" Doctoral Program (LPI–Paris), a Mind Science Foundation Grant to Lorenzo Ciccione and an ERC grant ("Neurosyntax") to Stanislas Dehaene.

Lorenzo Ciccione served as lead for data curation and writing–original draft, contributed equally to conceptualization, formal analysis, investigation, methodology, and visualization, and served in a supporting role for funding acquisition. Guillaume Dehaene contributed equally to methodology and writing–review and editing. Stanislas Dehaene served as lead for funding acquisition and supervision, contributed equally to conceptualization, formal analysis, investigation, methodology, visualization, and writing–review and editing, and served in a supporting role for data curation.

Correspondence concerning this article should be addressed to Lorenzo Ciccione, UNICOG Cognitive Neuroimaging Lab, CEA, INSERM, Université Paris-Saclay, NeuroSpin Center, CEA Saclay, NEUROSPIN - Bat. 145, Gif-sur-Yvette 91191, France. Email: lorenzo.ciccione@cri-paris.org

Every scientist regularly deals with outliers; i.e., anomalous observations or measurements that appear very different from the others. The dilemma is always the same: should we consider them the result of normal variability ("noise") inherent in the data, or exclude them from the main analysis, because they "arise suspicions that they were generated by a different mechanism" (Hawkins, 1980)? The answer is never straightforward and often depends on the data format, the scientific field, the number of observations and many other factors; as a consequence, several methods for outlier detection exist, with their advantages and disadvantages (Smiti, 2020). They include distribution-based methods (defining outliers as a function of their variation from a standard distribution), distance-based methods (which compute the distances among all items in the dataset, and consider as outliers those items that do not have close neighbors), and density and cluster-based approaches (which define outliers on the basis of their local density and their belonging to a distinct data cluster). Crucially, all of these methods depend on a threshold, a point beyond which an outlier is considered as such—and once a threshold has been fixed, they are only meant to detect outliers and do not provide explicit guidance on their inclusion or rejection from further analysis. A notable exception is represented by Bayesian approaches, which will be discussed later in the article. Interestingly, different graphical adaptations have also been proposed to facilitate the perceptual identification of outliers in graphs by human readers, through data visualization tools such as modifying the size, color, and opacity of different data points (Micallef et al., 2017).

One of the most intuitive (but still efficient) techniques to detect outliers is to plot all observations in a bivariate visual format, the scatterplot (Friendly & Denis, 2005), and to let a human viewer decide on the presence of outliers. Indeed, researchers in psychology consider scatterplots their elective tool in outlier detection (Orr et al., 1991). Alternatives, such as bar plots, hide the complexity of a dataset and may ultimately favor misleading conclusions about the data (Godau et al., 2016; Pastore et al., 2017). For example, scatterplots can be used to detect and reject response times (RTs) that are either too fast or too slow relative to the average value. They can also be useful to detect the existence of a secondary pattern of data that should be analyzed separately or in interaction (see Sunday et al., 2019 for an example).

These studies, however, raise an important and understudied question: are humans really capable of spotting outliers when a large dataset is displayed as a scatterplot? A growing body of research on "intuitive statistics" indicates that human adults are remarkably accurate at performing several different statistical tasks on scatterplots, such as linear trend judgment, extrapolation, and correlation estimation (e.g., Ciccione et al., 2022; Ciccione & Dehaene, 2021; Reimann et al., 2020; Rensink & Baldridge, 2010; Schulz et al., 2017). A recent study revealed that participants facing a scatterplot can perform a mental regression in a manner resembling a normative statistical model: they can judge the ascending or descending trend of a linear graph, and even estimate its slope, with a performance tightly correlated with the t value that a statistician would compute to evaluate the correlation in the graph (Ciccione & Dehaene, 2021).

The stimuli in those previous studies were always graphs without outliers, whose data points were normally distributed around the regression line. Only a few studies specifically investigated the role of outliers in graph-based tasks. One study found that human adults fail to fully reject outliers when asked to determine the Pearson r of the dataset (Bobko & Karren, 1979; Meyer et al., 1997). Similarly, correlation estimations are affected by outliers independently of the participants' statistical knowledge (Meyer & Shinar, 1992). Even when asked to adjust the trend on a scatterplot including outliers that are either extremely far or located at the boundaries of the main dataset, participants perform a linear regression that falls in-between a robust one (that excludes those outliers) and the line predicted by an ordinary least squares (OLS) algorithm (Correll & Heer, 2017; Liu et al., 2021). Taken together, these results suggest that, in the presence of clear and extreme outliers, participants are affected by them in their correlation judgments and regression estimations, although they assign them a lower weight than that of other observations in the dataset. In other words, participants attempt to reject outliers, but do not seem to be completely successful in doing so.

Another line of research on so-called "ensemble perception," the human ability to automatically encode summary statistics of the visual environment (for a review, see Whitney & Yamanashi Leib, 2018), found that, with stimuli other than graphs, humans do the exact opposite: they discard from their judgments all the items that considerably deviate from the other elements in the set (Epstein et al., 2020; Haberman & Whitney, 2010). This automatic filtering of outliers might indeed be highly beneficial in real-life contexts: avoiding deviant observations while focusing on the most representative information, allows to overcome our attentional limitations and to enhance our visual cognition (Alvarez, 2011).

Perceiving noisy graphical representations such as scatterplots might thus be a novel instance of ensemble perception (since humans manage to quickly and accurately extract a statistical trend from noise). At the same time, however, it does not seem robust to the presence of outliers, contrary to what the literature on ensemble coding would predict. Unfortunately, all past experimental investigations on outlier processing in graphs do not resolve this discrepancy. Indeed, previous studies share two fundamental limitations. First, they allowed participants to slowly inspect the scatterplot before providing any correlation judgment. Second, they always used outliers that diverged dramatically from the main distribution or that were located exclusively at its boundaries, without experimentally manipulating the strength of the outliers in terms of both their distance and their number. These experimental choices can surely be praised for their resemblance to ecological real-life situations: researchers usually take their time to inspect a graph and they often tend to reject only extreme outliers (Anscombe, 1960). However, they do not allow to characterize humans' spontaneous processing of outliers and their role in affecting intuitive statistics. Furthermore, they do not clearly separate outlier detection from outlier rejection, two processes that we suggest should be carefully distinguished—indeed, the above results suggest, but do not prove, that humans may detect the presence of outliers, and yet continue to be dragged toward them in their mental regression evaluations.

In the present series of three behavioral experiments, we aimed to provide an in-depth psychophysical investigation of the

perceptual processing of outliers in scatterplots. We tried to answer five open questions:

1) Do subjects spontaneously reject outliers when asked to perform a trend judgment or a regression estimation on a graph, without being told that there might be outliers? The aforementioned studies on ensemble perception (Whitney & Yamanashi Leib, 2018) found that outlier facial expressions (Haberman & Whitney, 2010) and oriented lines (Epstein et al., 2020) are spontaneously excluded when participants are asked to evaluate the average value of a set. We tested whether those findings extend to trend judgments and line fitting on scatterplots or whether, in this case, outlier items are automatically included.

2) Do the number of outliers and their distance from the main dataset modulate the bias that participants exhibit in estimating the slope or in judging the direction of the data's linear trend? Previous research (Bobko & Karren, 1979; Correll & Heer, 2017; Meyer & Shinar, 1992; Meyer et al., 1997) showed that correlation judgments and regression estimates are not robust to the presence of outliers, but this result could vary with the number and distance of the outliers. We thus measured if human performance in intuitive statistics is parametrically affected by those factors.

3) If outliers do bias participants' performance, is this bias modulated by the level of attention toward them? Across our three studies, we varied the level of attention to outliers by either not providing any information about their presence (Experiment 1: no attention); telling participants about their presence and inviting them to discard them in their judgments (Experiment 2: medium attention); or explicitly asking them to detect the presence of any outlier, on every trial, before estimating the line through the remaining data points (Experiment 3: high attention). There is very little prior research on this topic. Attention toward deviant stimuli has been shown to bias ensemble average estimations in the direction of the deviant item, but participants were never asked to discard outliers (de Fockert & Marchant, 2008). Our manipulation of participants' attention toward outliers thus provides a first test of the role of attention in outlier rejection.

4) How does outlier detection work? In Experiment 3, we asked participants to detect as fast as possible the presence (or absence) of any outlier in the dataset. In this manner, we could directly investigate the variables that affect outlier detection and, ultimately, to propose a model of how humans decide whether a given data point is an outlier or not.

5) If outliers are correctly detected, does this mean that they can also be rejected? In Experiment 3, we tried to disentangle outlier detection and rejection. On every trial, participants performed a task of outlier detection followed by slope estimation, thus allowing us to examine the contingencies between them. Participants might be well aware of the presence of outliers and the need to discard them, but still fail at doing so, thus suggesting that perceptually rejecting outliers is an ability impenetrable to cognition, as is the case for many visual phenomena (Stokes, 2013).

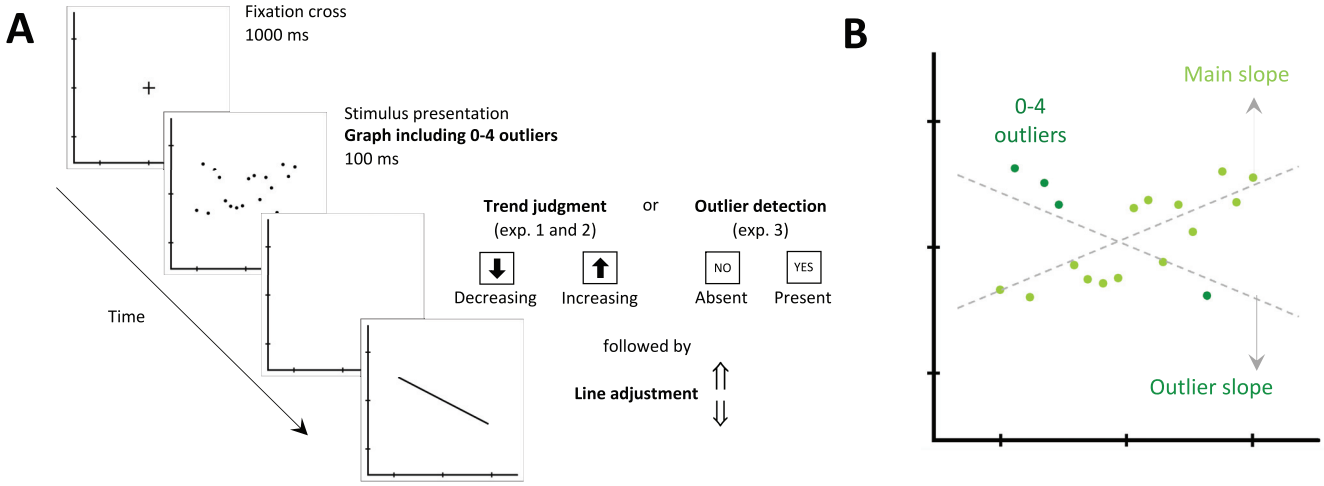
Method

Stimuli

All stimuli included two unlabeled lines denoting the x and y axes, which remained on screen for the duration of the experiment (Figure 1). Each line was marked with three small ticks at locations corresponding to the values 0, .5, and 1 (those numbers were arbitrary and not shown to participants). Within the area comprised by those two axes, the stimuli were scatterplots comprising 18 white dots on a black background. The x coordinates of the 18 points were fixed and separated by an equal distance on the x axis. Each stimulus was the graphical representation of a dataset generated on the basis of three experimental factors, whose values were combined in a full factorial design. First, we varied the slope of the line (the "main slope") around which the main datapoints (except outliers) were located; the main slope could take value: $-.5$, $-.25$, $+.25$, or $+.5$. Second, we independently varied the slope of the line around which the outliers were located; this "outliers' slope" could take value: $-.5$, $-.25$, $+.25$, or $+.5$. Third, we varied the number of outliers (n = 0, 1, 2, 3, or 4). In detail, the stimulus generation algorithm worked as follows. First, the y coordinates of all points were determined according to the following equation: $y_i = \text{main\_slope} \times x_i + \varepsilon_i$, where the x_i are 18 numbers equally spaced between 0 and 1, and the ε_i are random numbers independently drawn from a normal distribution centered on zero and with standard deviation of .1. Afterward, the desired number of outlier points (0, 1, 2, 3, or 4) were selected at random among all points in the dataset, excepting the six central ones, and their y coordinates were changed according to the following equation: $y_i = \text{outliers\_slope} \times x_i + \varepsilon_i$, again with $\varepsilon_i \in N(0, .1)$. Because of the added noise, the OLS regression slope of the nonoutliers dots could depart slightly from the prescribed one ("main slope"). To compensate for this, a small linear component was added to the main datapoints, calculated such that their final slope corresponded precisely to the prescribed one, and always passed through the center of the screen (see Figure 1B for an example with a main slope of .5, an outliers' slope of $-.5$ and four outliers. Examples of stimuli are provided in online supplementary materials). When the main slope was identical to the outlier slope, all data points were generated around a single slope, thus resulting in no outlier being presented (and such condition was considered equivalent to the one with 0 prescribed outliers).

We generated outliers using a secondary process (namely another regression line with the outlier slope) because it offered a means to finely control their average distance from the main dataset, while still avoiding to impose an exact location to them. A different choice would have been to manipulate the distance factor by using different standard deviation distances from the main regression line but, in this case, all outliers would have had, for a given distance condition, exactly the same deviance, likely making the stimuli easily recognizable over trials.

Figure 1
Experimental Design



Note. (A) Example trial. On each trial, participants were presented with a scatterplot and asked to judge, in Experiment 1 and 2, if its trend was ascending or descending; or, in Experiment 3, if there were any outliers. Immediately after their response, they had to adjust the slope of a line on screen by moving their finger on a trackpad, in order to provide an estimation of the regression line underlying the noisy scatterplot. In Experiment 1, participants were not informed of the presence of outliers. Experiment 2 differed from Experiment 1 only in that participants were informed that some outliers could be present, and were asked to try to ignore them in their judgments. Experiment 3 further emphasized outliers by first asking for explicit outlier detection before the slope adjustment task. (B) Illustration of the stimulus generation process. In each scatterplot, the majority of dots were noisy samples around a line with a main slope of either 0.5, 0.25, −0.25, or −0.5. Between zero and four dots were outliers (in this example, 4) generated as noisy samples around another line, whose slope could also take the values 0.5, 0.25, −0.25 or −0.5. Different colors are used for illustrations purposes only: outliers were not signaled in any way, since all stimuli were white dots on a black background. See the online article for the color version of this figure.

Participants

Thirty participants (10 per experiment) were recruited (age: 26.2 ± 2.1, 16 females, 14 males). The sample size was the same as in previous studies that used the same type of stimuli (Ciccione & Dehaene, 2021). We computed a power analysis (using the G*power software; see Faul et al., 2007) for our main ANOVAs of error rates and response bias with three within subjects' factors (slope, noise, and number of points), three groups, and five repeated measures per subject and condition. We used an α level of .05, a power of .9, an effect size (partial η^2) of .1 (i.e., the smallest effect size found in our previous studies of graph perception), and a conservative expected correlation between repeated measures of .1 (in order to account for a possible large variability in responses across trials). This power analysis resulted in a recommended sample size of 27 participants, which we rounded to 30. All participants had normal or corrected-to-normal vision, no medical history of epilepsy, were right-handed, and did not take psychoactive drugs. The experiment was advertised through the mailing list of the first author's university. In order to ensure the homogeneity of the sample in terms of participants' cultural background, only participants with at least a master's degree were recruited. They all signed an informed consent and were paid 5 euros for their participation. The experimental sessions lasted approximately 30 minutes and were approved by the local ethical committee (under the reference CER-Paris-Saclay-2019-061). All experimental sessions took place in 2021. One participant was excluded from Experiment 2 analyses since he failed to perform the task appropriately (his performance was at chance level).

It is worth noting that, given the recruitment method and the desire to maintain a homogenous sample, the results might not generalize to younger or older populations and/or to people with a lower or higher expertise and familiarity with graphical representations.

Experimental Procedure

Participants were invited to sit on a fixed chair with their head at a distance of 50 cm from the screen. Each experimental session was divided into five blocks of 80 trials; the duration of each block was ∼4 minutes. After each block, participants could take a short break. Before starting the actual experiment, 25 practice trials were run under the researcher's supervision, in order to control for the correct execution of the task (i.e., maintaining the correct distance from the screen, correctly placing their hand and fingers; familiarizing with the rapid presentation of the stimuli. No feedback on performance was provided). On each trial, as illustrated in Figure 1A, a fixation cross first appeared for 1,000 ms, immediately followed by a scatterplot flashed for 100 ms. The stimuli were flashed in order to promote spontaneous and fast responses and thus to avoid any possible explicit strategy, calculation, or complex eye movement patterns. The experimental procedure varied depending on the experiment.

Experiment 1

No information concerning the presence of outliers was given to participants. They were merely asked to respond (as fast and accurately as possible) by pressing with their left-hand ring finger on a

key (signaled with a ⇓ sticker) if they thought that the trend in the scatterplot was decreasing or, conversely, to press with their left-hand index finger on another key (signaled with a ⇑ sticker) if they thought that the trend in the scatterplot was increasing. Immediately after this first response, an adjustable line appeared in the middle of the screen. The line was initially horizontal, but participants were asked to adjust it as accurately as possible by moving their right-hand index finger on the computer trackpad. The center of the line was kept fixed, 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 the finger displacement; moving the finger up tilted the line in the counterclockwise direction, whereas moving the finger down tilted it in the clockwise direction. For this second task, participants were invited to respond independently of their first trend judgment: they were explicitly told that they could orient the line in a direction opposite to their trend judgment, if they thought that they had made a mistake in the first task. When the adjustment was completed, they pressed the trackpad in order to confirm their answer and move to the next trial, which was preceded by a 1-s fixation cross.

Experiment 2

Participants were asked to perform the exact same task as in Experiment 1. The only difference consisted in the information given to them before starting the experimental session: participants were informed that one or more outliers, defined as points outside the main dataset, could be present in some trials. They were invited to try to exclude such outliers from their answers, and thus to perform both tasks of trend judgment and slope adjustment only on the main dataset.

Experiment 3

As in Experiment 2, participants were informed that one or more outliers could be present in some trials. They were asked to detect them, as fast and accurately as possible, by pressing with their left-hand ring finger on a key (signaled with a "NO" sticker) if they thought that the scatterplot did not include any outliers or, conversely, to press with their left-hand index finger on another key (signaled with a "YES" sticker) if they thought that the scatterplot included one or more outliers. Immediately after this detection response, they moved to the slope adjustment task, identical to Experiment 2, with the explicit instruction to try to estimate the slope of the main dataset only and, thus, to reject outliers.

Transparency and Openness

All data and scripts for the analyses are available on the Open Science Framework at: https://osf.io/bfjsw/.

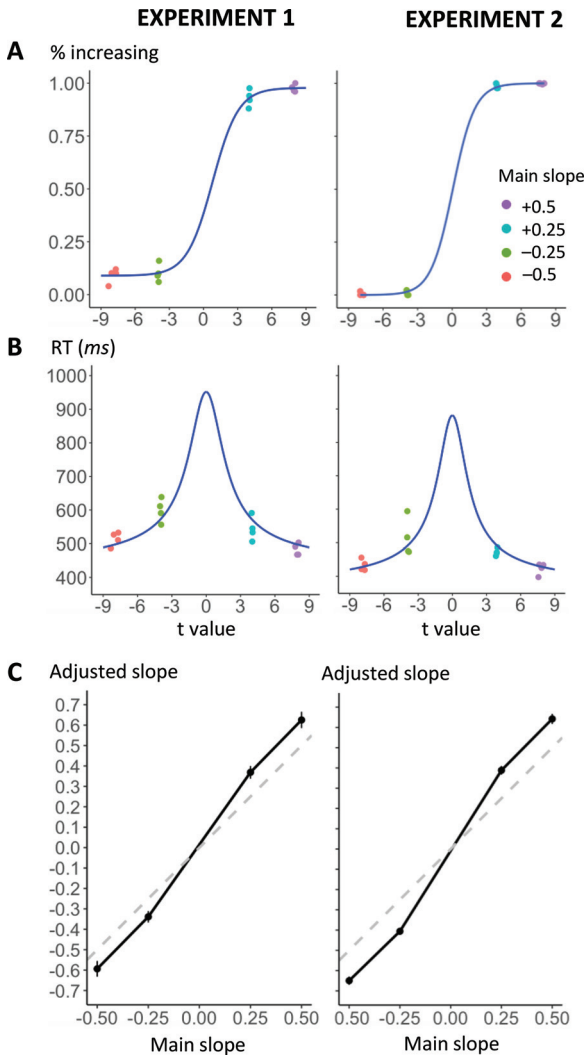
Results

Performance in Trend Judgment and Line Adjustment in Graphs Without Outliers

First, in an attempt to replicate our previous work (Ciccione & Dehaene, 2021), we analyzed participants' trend judgment performance in the absence of outliers. Figure 2 (top) shows the percentage of trials classified as "increasing" as a function of the main slope and of the *t* value associated with the scatterplot linear regression. The *t* value we used is the one that a statistician would calculate to test for the statistical significance of a positive or negative trend in a dataset and it has been previously shown to predict human mental regression judgments (Ciccione & Dehaene, 2021). As clear from the figures, in both Experiments 1 and 2, participants' responses could be modeled as a sigmoid function of such

Figure 2

Performance in Trend Judgment (A, B) and Line Adjustment (C) in Graphs Without Outliers



Note. In both experiments, the percentage of "increasing" responses (A) and the response times (B) vary systematically with the *t* value associated to the Pearson coefficient of correlation. The blue (black) lines in the middle row indicate the response times predicted by a simple accumulation-of-evidence model (Gold and Shadlen, 2002). The plots in (C) show the slopes reported by participants (black lines) and predicted by ordinary least squares (OLS) regression (dashed gray lines), which corresponds to the process by which the scatterplots were generated. Participants responded with slopes exceeding those predicted by OLS, in agreement with the use of Deming regression. All of these results replicate our previous findings with similar mental regression tasks (Ciccione & Dehaene, 2021). See the online article for the color version of this figure.

t value. Both the sigmoidal shape of their response rates (Figure 2, top) and the distance effect in their response times (i.e., slower responses for stimuli with a *t*-value closer to zero; Figure 2, middle) could be jointly predicted by a classical decision-making model, which assumes a noisy accumulation of evidence toward a decision bound (Gold & Shadlen, 2002). In Figure 2, the blue (black) lines show the performance predicted by that model. These results replicate previous findings on human mental regression (Ciccione & Dehaene, 2021; see that article for modeling details).

Influence of Outliers on Accuracy in the Trend Judgment Task

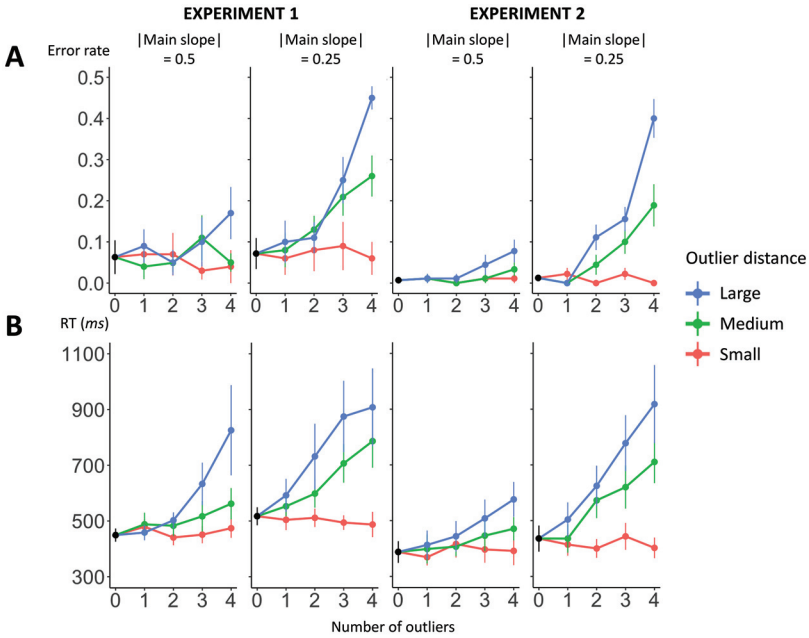
We then looked at participants' performance in the same trend judgment task ("ascending or descending?") when outliers were present in the stimuli. Figure 3 shows the results from both Experiments 1 and 2, which closely resembled each other. The average error rates for stimuli without outliers are simply indicated as a reference (the black dots). The top row indicates the error rate as a function of the number of outliers as well as two other driving variables: the absolute value of the main slope of the scatterplot (steep: .5; or shallow: .25), and the outliers' distance, quantified as the absolute difference between the outliers' slope and the main slope. For a main slope of .5, the available outliers' distances were 1, .75, and .25, which for simplicity are referred to, respectively, as *large*, *medium*, and *small*. Similarly, for a main slope of .25, the available outliers' distances were .75, .5, and .25, which are again referred to with the same labels.

As we can see, in both experiments, the error rate increased as a function of the number of outliers and their distance from the main dataset, indicating that the participants' responses were attracted toward the outliers. To test for the significance of these observations, we conducted an ANOVA on participants' error rates with the main slope, the outliers' distance and the number of outliers as within-subjects' factors and the experiment number (1 or 2) as a between-subjects' factor. While the experiment number had no significant main effect nor any interaction (all related p values $> .1$), all other factors had main effects (main slope: $F(1, 17) = 101.15$, partial $\eta^2 = .86$, $p < .001$; outliers' distance: $F[1.5, 25.8] = 90.97$, partial $\eta^2 = .84$, $p < .001$; number of outliers: $F[2.5, 42.9] = 39.55$, partial $\eta^2 = .7$, $p < .001$) and interaction effects (main slope and outliers' distance: $F[1.6, 27.4] = 18.78$, partial $\eta^2 = .52$, $p < .001$; main slope and number of outliers: $F[2.3, 39] = 25.6$, partial $\eta^2 = .6$, $p < .001$; outliers' distance and number of outliers: $F[3.2, 54.1] = 19.12$, partial $\eta^2 = .53$, $p < .001$; triple interaction of main slope, outliers' distance and number of outliers: $F[3.9, 66.9] = 6.54$, partial $\eta^2 = .28$, $p < .001$).

Figure 3 clarifies the meaning of those interactions. First, error rates generally increase with the number of outliers, but more so when the main slope is shallow (.25), thus rendering the main decision more difficult, than when the main slope is steep (.5).

Figure 3

Influence of Outliers on Performance in the Trend Judgment Task ("Is the Graph Ascending or Descending?")



Note. Results are plotted as a function of the number of outliers, separately for graphs with steep (0.5) or shallow (0.25) main slopes. Both error rates (A) and response times (B) increase as a function of the number of outliers, as well as of the distance of the outlier slope from the main slope. When the graph has a shallower main slope (0.25), thus rendering the task more difficult, the influence of outliers becomes correspondingly larger. Error bars indicate one standard error of the mean across subjects. See the online article for the color version of this figure.

Second, similarly, for the same number of outliers, their impact is larger when their distance to the main dataset is larger; i.e., when they deviate more from the main regression line. Those findings make sense: essentially, the more numerous the outliers, and the more they push toward a line with a different orientation from the main one, the more likely participants are to make an error. Indeed, it is worth noting that the small outliers' distance condition (red [black] lines), with a main slope of .25, was the only experimental condition in which the outliers' slope was steeper than the main slope: in this situation, outliers were not expected to make the trend judgment harder to perform, since they made the overall trend of the graph steeper. Indeed, the error rates in these conditions did not increase as a function of the number of outliers (the red [black] lines are essentially flat): this was confirmed by a nonsignificant main effect of the number of outliers in two ANOVAs restricted to those conditions (Experiment 1: $F[1.1, 9.7] = .28$, partial $\eta^2 = .03$, $p = .62$; Experiment 2: $F(3, 24) = 1.73$, partial $\eta^2 = .18$, $p = .19$).

Influence of Outliers on Response Times in the Trend Judgment Task

Response times in the trend judgment task for Experiments 1 and 2 are plotted in Figure 3 (bottom). Response times behaved in parallel to error rates, thus indicating the absence of a speed/accuracy tradeoff. They increased as a function of the number of outliers as well as of the distance of the outliers' slope from the main slope. The same ANOVA as above, now on median response times, again revealed no main effect or interactions involving the experiment factor (all related p values $> .1$). It also indicated that all within-subject factors had a significant main effect (main slope: $F(1, 17) = 53.67$, partial $\eta^2 = .76$, $p < .001$; outliers' distance: $F[1.1, 18] = 26.88$, partial $\eta^2 = .61$, $p < .001$; number of outliers: $F[1.6, 26.8] = 20.02$, partial $\eta^2 = .54$, $p < .001$) and entered into significant interactions (main slope and outliers' distance: $F[1.4, 23.3] = 16.87$, partial $\eta^2 = .5$, $p < .001$; main slope and number of outliers: $F[2.4, 40.3] = 6.45$, partial $\eta^2 = .28$, $p < .01$; outliers' distance and number of outliers: $F[2.4, 39.9] = 15.04$, partial $\eta^2 = .47$, $p < .001$; no triple interaction of the within-subjects factors was found). Again, those interaction effects are easily observable in Figure 3: response times increased significantly faster with the number of outliers as the distance of the outliers increases, and also as the main slope gets shallower. Like for error rates, the experimental condition in which the outliers' slope was steeper than the main one resulted in no increase of response times (as evident from the essentially flat red [black] lines in the plots for a main slope of .25), which was confirmed by two ANOVAs restricted to those conditions (Experiment 1: $F[2.6, 23.4] = .7$, partial $\eta^2 = .07$, $p = .54$; Experiment 2: $F[2, 15.7] = .37$, partial $\eta^2 = .04$, $p = .7$).

Lastly, as we can see from the response time plots for a main slope of .25, we found that the presence of a single outlier, at a large enough distance from the main dataset (blue [light gray] lines), induced a substantial increase in response times. Indeed, a paired t-test on participants' response times from both experiments revealed a significantly slower median response time in the presence of one outlier than in the absence of outliers ($t(18) = 2.78$, $p < .01$; respectively 550 versus 478 ms).

One could argue that a greater number of outliers simply made the overall slope of the dataset closer to zero, thus making trend judgment more difficult. Could participants' slower response times be explained by changes in slope rather than by the number of outliers? To test for this, we performed a multiple linear regression on response times with both the number of outliers and the absolute Deming slope as predictors. Note that we used Deming regression, which minimizes the orthogonal distance of the points to the fit, instead of classic OLS regression, because previous evidence (Ciccione & Dehaene, 2021) showed that humans use this procedure in their intuitive mental regressions (this is discussed further below). We found that both were significant ($\beta_{\text{number of outliers}} = 20.7$ ms/outlier, $p < .0001$; $\beta_{\text{absolute Deming slope}} = -822.5$, $p < .0001$). We also computed a linear regression on the residuals of the response times as a function of the absolute Deming slope and found that the number of outliers was still a significant predictor ($\beta = 19.2$ ms/outlier, $p < .0001$). Thus, the results confirm that outliers influenced response times over and above their indirect effect on the overall trend, with a cost of $\sim$20 ms per outlier.

Overall, performance in the trend judgment task indicated that, regardless of the instructions to exclude outliers, participants were always strongly influenced by them, especially when (a) they were more numerous; (b) their deviation was large; and (c) the decision was difficult because the main slope was shallow.

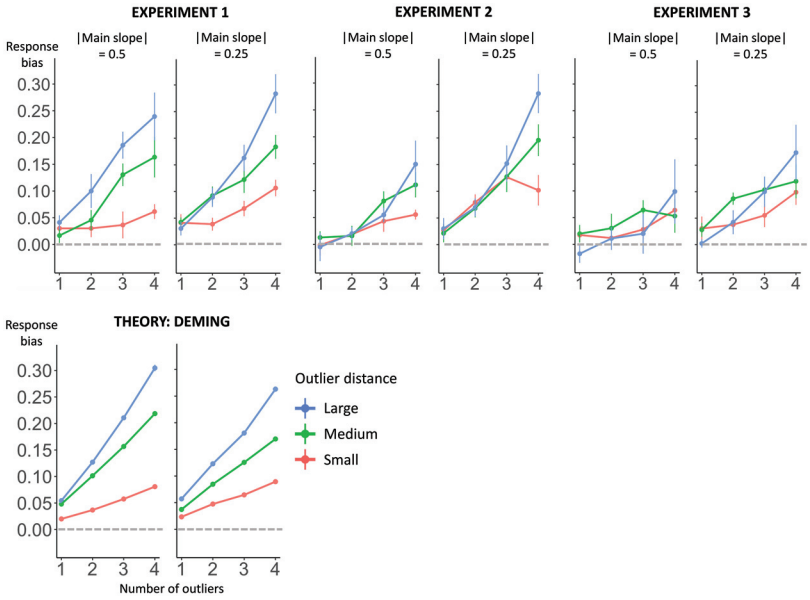
Influence of Outliers on the Line Adjustment Task

The second task, which was run in Experiments 1, 2, and 3, consisted of a slope adjustment: participants were asked to adjust the line in order to best fit the scatterplot. As explained in the methods section, the three experiments differed only in terms of the induced level of attention about the presence of outliers in the stimuli: in Experiment 1, no information about outliers was given; in Experiment 2, participants were invited to exclude them in both their trend judgment and slope adjustment; in Experiment 3, they were explicitly invited to concentrate on them and detect their presence (or absence) before performing the slope adjustment task (after rejecting them).

We first examined the slope estimated by participants in the absence of outliers (Figure 2C shows the results for Experiments 1 and 2, which were similar to Experiment 3). Confirming previous evidence (Ciccione & Dehaene, 2021), we found that the estimated slopes closely tracked the actual slopes of the graphs, but were steeper than the ones predicted by a classic ordinary least squares (OLS) regression (the gray dashed lines in Figure 2C). Their values were compatible with the minimization of the orthogonal distance of the points to the best-fitting line, a procedure known as Deming regression.

For each subject and each experimental condition, we then evaluated the impact of outliers relative to this no-outlier baseline. To this aim, we calculated "response bias" as the difference between the median slope that they reported in the presence of outliers and in their absence. For visualization and analysis' purposes, the sign of this difference was flipped such that a positive value always indicated attraction toward the outliers (in practice, this meant that we flipped the sign for all stimuli with an outliers' slope lower than the main slope). Figure 4 shows the mean response bias as a function of experiment, main slope, number of

Figure 4
Influence of Outliers on the Adjusted Slope in the Line Adjustment Task



Note. Top panels: results of Experiments 1, 2, 3, separately for graphs with a steep (0.5) and a shallow main slope (0.25). Response bias was calculated as the average difference between the slope reported in the presence of a certain number of outliers (x axis) minus the slope reported in the absence of any outliers. Data were flipped such that a positive value always indicates attraction towards the outliers. Across experiments, the bias decreases, thus suggesting an improved rejection of outliers. Error bars indicate one standard error of the mean across subjects. Bottom panels: theoretical predictions of Deming regression. Response bias was calculated as the average difference between the overall Deming slope (including outliers) and the Deming slope in the absence of any outliers. The response bias in Experiment 1 was almost identical to the response bias of Deming regression, thus confirming that participants, when not informed about the presence of outliers, performed a Deming regression on the entire dataset. See the online article for the color version of this figure.

outliers, and outliers' distance from the main dataset. We can see that the outlier-induced bias increased with the number of outliers, but did so faster for a large outliers' distance, and more so in Experiment 1 than in Experiment 2 or, a fortiori, Experiment 3. We confirmed these observations through a repeated measures ANOVA on participants' median bias with experiment number as between-subjects factor and main slope, number of outliers, and outliers' distance as within-subjects factors. All of the latter had a significant main effect (main slope: $F(1, 26) = 43.35$, partial $\eta^2 = .63$, $p < .001$; number of outliers: $F[1.57, 40.94] = 82.72$, partial $\eta^2 = .76$, $p < .001$; outliers' distance: $F[1.48, 38.53] = 22.08$, partial $\eta^2 = .46$, $p < .001$). Although the main effect of *experiment* was close to significance ($F(2, 26) = 3.20$, partial $\eta^2 = .2$, $p = .06$), it entered in a significant interaction with both the main slope ($F(2, 26) = 4.99$, partial $\eta^2 = .28$, $p = .01$) and the outliers' distance ($F[2.96, 38.53] = 5.76$, partial $\eta^2 = .31$, $p < .01$). Indeed, as we can see from Figure 4, the outlier-induced bias decreased across experiments, as the level of attention to outliers increased, and this effect was more pronounced for a larger number of outliers and for larger outliers' distances. It is worth noting that the number of outliers had also a significant interaction with both the main slope ($F[2.17, 56.35] = 3.9$, partial $\eta^2 = .13$, $p = .02$) and the outliers' distance ($F[4.15, 107.81] = 12.36$, partial $\eta^2 = .32$, $p < .001$). Thus, the results of the line adjustment task (Figure 4) closely paralleled those of the trend judgment task (Figure 3).

If, as we suggest, uninformed participants did not spontaneously reject outliers, but included them in their regression estimates, then their response bias should be predictable by a global regression performed on the entire dataset. To test this idea, we examined whether the response bias from participants of Experiment 1 (i.e., those who received no information about the presence of outliers) mirrored the theoretical predictions of Deming regression. As with the actual data, we first computed the response bias as the difference between the slope predicted when the regression was applied to the entire dataset, and when it was applied to a dataset without outliers. Figure 4 (bottom) shows the predicted biases for each experimental condition, plotted in the same way as the human data. Those predictions quantitatively match the observed data (linear regression between predicted and observed, $R^2 = .91$, slope = $1.02 \pm .07$, intercept = .01). In particular, Deming regression predicts that bias should increase with the number of outliers and with their distance from the main dataset, exactly as in human data.

Performance in Outlier Detection

On every trial of Experiment 3, participants first performed an outlier detection task: immediately after the flashing of the scatterplot, they had to decide whether they had seen at least one outlier or not, by pressing one of two response keys as fast and accurately as possible. This experimental procedure allowed us to directly investigate whether and how humans detect the presence of outliers. Figure 5 shows the percentage of "yes" responses as a function of the main slope, the number of outliers and their distance. The results indicate that false alarms were quite high (40–50% of trials without outliers), but that correct detection increased as a function of the number of outliers, especially for large and medium outliers' distances. Those observations were confirmed by an ANOVA on the percentage of "yes" responses with the above factors as within-subjects' factors (to obtain a full factorial design, we excluded the conditions with 0 outliers, which are presented in Figure 5 only for reference). There was a main effect of both the number of outliers (F[1.90, 17.13] = 11.52, partial η^2 = .56, $p <$ .001) and their distance (F[1.59, 14.28] = 52.61, partial η^2 = .85, $p <$ .001). The main slope had no main effect (p = .48) but entered in a significant interaction with the outliers' distance (F[1.79, 16.13] = 11.32, partial η^2 = .56, p = .001): in fact, as clear from Figure 5, for a steeper main slope of .5, the difference in correct detections between the three outliers' distances was more pronounced than for a main slope of .25.

We ran a similar ANOVA on participants' median response times for correct detections and found only a significant main effect of outliers' distance (F[1.13, 10.17] = 5.99, partial η^2 = .4, p = .03) and its interaction with the main slope (F[1.81, 16.25] = 6.09, partial η^2 = .4, p = .01).

Formulating and Testing a Theory of Outlier Detection and Rejection

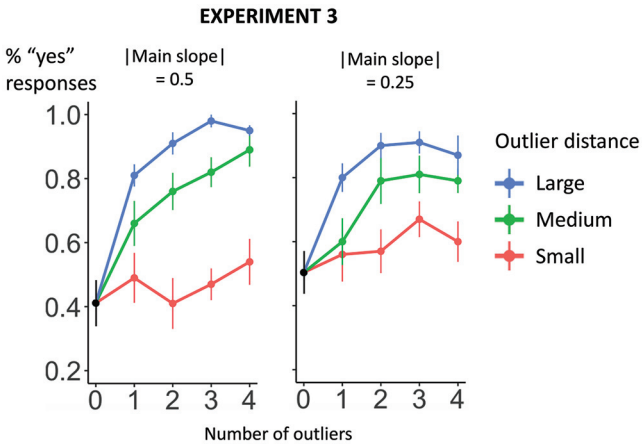
On what basis do participants decide on the presence of outliers? We formulated the hypothesis that, like a statistician, they might base their judgments on an estimate of how much a given data point departs from the rest of the cloud. A simple way of measuring such a departure is to compute a z-score for each point; i.e., a fraction with the numerator equal to the distance of that point to the regression line, and the denominator equal to the standard deviation of such distances. Such a z-score evaluates to what extent the observed data point is out-of-distribution compared to the other ones.

The specific model we propose is shown in Figure 6A. Two choices were made. First, since we know from the present and past research (Ciccione & Dehaene, 2021) that participants compute their mental regressions by minimizing the perpendicular distance of the points from the best-fitting line, as in Deming regression, rather than the vertical distance as in OLS regression, we computed the Deming regression of each scatterplot and postulated that, for the numerator, participants use the perpendicular distance to that line. Second, for the denominator, since our graphs all had the same noise level (SD = .1), we postulated that subjects could pool their noise estimates across trials and eventually converge to a fixed value. Note that this hypothesis may be revised in a different experimental setting—for instance if participants saw a single graph, or if the noise level varied across trials; then their estimate could be based on the observed graph. Here, however, we obtained a better account by postulating a fixed value of the denominator (as confirmed by a model comparison described later in this section).

In the end, we therefore calculated, for each point, a z-score equal to its perpendicular distance to the regression line divided by .1 (Figure 6A). Our hypothesis predicts that this value is the decision variable on the basis of which participants decide whether that point is an outlier. Since they had to decide whether any outlier was present, the percentage of "yes" responses in outlier detection should be a logistic function of the maximum z-score over all 18 data points. Figure 6B shows the corresponding psychophysical curve (for visualization and analysis' purposes, the responses were binned according to the highest z-score). We ran a multiple logistic regression on all participants' responses with two regressors: the highest z-score and the actual prescribed number of outliers; we found that the former was an excellent predictor of "yes" responses (β = 1.91, $p <$ .0001), better than the actual prescribed number of outliers (β = .1, $p <$ .0001). Indeed, as we can see from Figure 6B, when the highest z-score was low ($\sim$.8), the proportion of "yes" responses dropped to 15%, lower than the average rates of false alarms of 48% on trials where prescribed outliers were genuinely absent (Figure 5). Conversely, at the opposite extreme, when the highest z-score exceeded about 3, the detection rate was close to 100%, higher than the average values of 65% when a single outlier was actually present (Figure 5).

One could rightfully argue that the highest z-score in the dataset does not take into account the number of other outliers. To focus on the simplest cases, we thus restricted our logistic regression to stimuli with either no prescribed outlier, or with a single prescribed outlier—and in both cases, we found that the highest z-score was still a significant predictor of the percentage of "yes" responses (respectively: $\beta_{0\_outliers}$ = 1.88, $p <$ .0001; $\beta_{1\_outlier}$ = 2.22, $p <$ .0001). Figure 6B makes it clear that a single function of the z-score provided an excellent account of the outlier detection responses, regardless of the actual prescribed number of outliers.

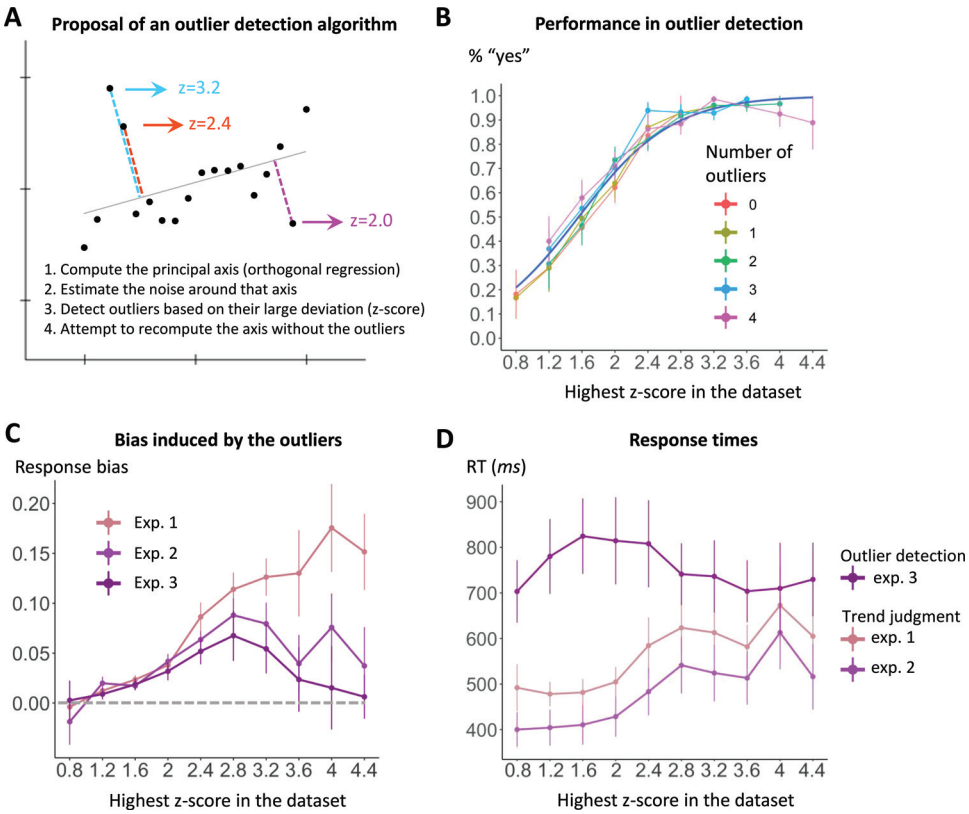
Figure 5
Performance in Outlier Detection in Experiment 3



Note. The percentage of trials in which participants reported seeing at least one outlier is plotted as a function of the true (i.e., "prescribed") number of outliers (0–4). This percentage increases as a function of the number of outliers, as well as their distance between their slope and the main slope. Error bars indicate one standard error of the mean across subjects. See the online article for the color version of this figure.

Figure 6

Participants May Detect Outliers by Computing the Significance of Their Deviation From the Principal Axis



Note. (A) example of a scatterplot with three outliers and proposal of an outlier detection algorithm. The outliers can be detected by calculating their individual *z*-scores, computed as their distance to the Deming regression line (i.e., the principal axis), divided by an estimate of the standard deviation of those distances. Outliers tend to have large *z* values. (B) the percentage of trials in which the subjects reported seeing at least one outlier (detection task in Experiment 3) is well predicted by the highest *z*-score in the stimulus graph, regardless of the prescribed number of outliers. Crucially, the highest *z*-score is a better predictor of outlier detection than the prescribed number of outliers, (shown in Figure 5). (C) the bias in the slope adjustment task also varies as a function of the highest *z*-score in the dataset. In Experiment 1, where participants were not told about outliers, the increase is essentially monotonic. In Experiments 2 and 3, the bias starts decreasing when the highest *z*-score exceeds ~2.8. In Experiment 3, the bias returns to zero for larger *z*-scores, indicating that extreme outliers can be rejected when explicitly instructed. (D) response times in trend judgment (Experiments 1 and 2) increase as a function of the highest *z*-score in the dataset; response times in outlier detection (Experiment 3) peak for highest *z*-scores around 2, where the model predicts the presence or absence of outliers to be most ambiguous. Error bars indicate one standard error of the mean across subjects. See the online article for the color version of this figure.

We tested several alternative ways of computing the *z*-scores. First, the distances (numerator) could be computed using the regression of all points (as we did) or the regression restricted to the main dataset. Second, they could be based on the perpendicular distance to the Deming regression, or the vertical distance to the OLS fit. Third, the standard deviation (denominator) could use the prescribed standard deviation of the distances (.1) or the actual standard deviation, measured from the specific graph. We modeled the logistic regressions of the percentage of "yes" responses as a function of the highest *z*-score calculated through all eight combinations of those three parameters and found that the model with the significantly smallest Akaike Information Criterion (AIC), thus the one more plausible to be correct (Akaike, 1998), was the above-described model.

Given that the highest *z*-score accounted well for outlier detection in Experiment 3, we next examined whether the same variable also predicted the capacity for outlier rejection; i.e., the influence of outliers on mental regression slopes. To this end, we went back to Experiments 1, 2, and 3, and plotted the participants' response bias in the line adjustment task as a function of the highest *z*-score in the stimulus graph, separately for each experiment (Figure 6C). Interestingly, for Experiment 1, the response bias increased monotonously as a function of the highest *z*-score (R^2 = .36, $F(1, 98) = 57.21$, $p <$ .001): with no information concerning the presence of outliers,

participants included them in their estimations, and the greater their deviance, the higher the bias they induced. However, for Experiments 2 and 3, in which participants were explicitly asked to reject outliers, a similar increase in response bias was seen only up to a highest z-score of $\sim$2.8, after which the bias started to decrease. Indeed, in Experiment 3, which required an explicit outlier detection on each trial, the bias was statistically indistinguishable from zero for z-scores higher than 3.6 (mean bias = .001; t-test on all responses against zero: $t(173) = .09$, $p = .93$).

These observations were confirmed by a repeated-measures ANOVA on the outlier-induced bias with experiment (1, 2, or 3) as between-subjects factor and the highest z-score in the dataset as within-subjects factor: both had a significant main effect (*Experiment*: $F(2, 26) = 4.94$, partial $\eta^2 = .28$, $p = .02$; highest z-score: $F[3.17, 82.53] = 8.56$, partial $\eta^2 = .25$, $p < .0001$) and entered into a significant interaction with each other ($F[6.35, 82.53] = 2.57$, partial $\eta^2 = .17$, $p = .02$). Crucially, the main effect of the experiment and its interaction with the highest z-score vanished when the ANOVA was computed only on stimuli with a highest z-score limited to values at or below 2.4 (both p values $> .47$).

In summary, the data in Figure 6 suggests the existence of two ranges. For highest z-scores below roughly 2.4, participants miss many of the outliers, while their influence on regression responses increase with z; and for highest z-scores above that value, outlier detection approaches 100%, and their influence on mental regression starts to decrease—but only if subjects are told to reject them.

This conclusion seems to suggest that, on average, outlier rejection closely parallels outlier detection. However, this was not true on a single-trial basis. We restricted the analysis to those trials of Experiment 3 in which (a) a single outlier was prescribed; (b) that point had the highest z-score; and (c) the participant responded that he had detected an outlier (most likely the prescribed one). On such trials, if outlier detection automatically led to outlier rejection, there should be no outlier-induced bias on the participants' slope estimates. This was true for scatterplots with one prescribed outlier with a z-score higher than 2 ($t(84) = -.31$, $p = .62$) but not for scatterplots with one prescribed outlier with a z-score at or below 2: for these stimuli, the bias was still significantly higher than zero ($t(49) = 2.75$, $p < .01$). This finding shows that participants could remain influenced even by outliers that they have detected.

Lastly, we looked at whether the response times could also be predicted by the z-score of the datapoints (Figure 6D). First, we considered the trend judgment task used in Experiments 1 and 2, where we previously found that RT increased with the prescribed number of outliers, and examined whether it could be explained by the actual number of outliers. To estimate the latter, we calculated, for each graph, the number of outliers passing a threshold of $z > 2$, and we included it as a predictor in a multiple regression on response times, together with the absolute Deming slope and the absolute main slope of the dataset. All predictors were significant ($\beta_{\text{number of outliers higher than z=2}} = 25.3$ ms/outlier, $p < .0001$; $\beta_{\text{absolute Deming slope}} = -1185.3$, $p < .0001$; $\beta_{\text{main slope}} = 579.2$, $p < .0001$). We then calculated the residuals of the regression with the two mentioned slopes as predictors and computed a linear regression on such residuals as a function of the number of outliers with a z-score higher than 2, finding it was still a significant predictor ($\beta = 13.9$, $p < .01$). Crucially, such a linear regression had an AIC of 112,289, which was significantly smaller than the one calculated on the residuals as a function of the prescribed number of outliers (AIC = 112,456, $\Delta_{\text{AIC}} = 167$, $p < .0001$), suggesting once more that the z-score of the datapoints was a better predictor of participants' performance than the prescribed number of outliers. This is evident when comparing Figure 5 to Figure 6B: if the prescribed number of outliers is taken into account (Figure 5), outliers are falsely detected at a very high rate ($\sim$40–50% when no outliers were present); however, when the actual distance of those outliers is considered (Figure 6B), the false detection rate turns out to be much lower ($\sim$20–30% for trials with a low z-score).

Next, we considered the response times in outlier detection (Experiment 3). Our model predicts that participants take that decision by evaluating whether any point has a z-score above a threshold value, close to $z = 2$. Thus, the decision variable should be the difference between the highest score and this threshold, and response times should be increasingly slower as this difference approaches zero. To test this prediction, for each graph, we calculated the absolute distance between its highest z-score and 2, and we used such value as a predictor in a linear regression of response times. The effect was significant ($\beta = -70.8$, $p < .0001$), and a plot of RTs indicated that indeed, RTs decreased with the distance from the putative decision boundary (Figure 6D).

Comparing Human Performance With an Optimal Bayesian Model

As explained in the introduction, formal methods of outlier detection share two fundamental aspects: they possess a threshold beyond which a datapoint is dichotomously considered an outlier or not, and they do not provide any explicit indication on whether the outlier should be included or excluded from the analysis—and thus do not directly speak to our data, which are primarily about how participants' regression estimates vary in the presence of outliers, and of instructions to reject them.

An exception is given by Bayesian approaches, which compute the posterior probability that each observation is an outlier; such probability can be seen as the "weight" that each item has in the regression (a lower probability/weight has a smaller influence on the regression). How does this approach perform in comparison with our participants? In order to answer this question, we computed, for each trial used in our experiments, the posterior probability of each item being an outlier, as formalized by Chaloner and Brant (1988). Then, for each such trial, we ran 1,000 iterations, in which the points in the dataset were excluded depending on their probability to be an outlier (e.g., a point with a probability of .8 being an outlier, was excluded, on average, 80% of the times). We then calculated the Deming regression slope of each iteration (i.e., on the items that, on that occasion, were not considered outliers) and took the median of the 1,000 iterations. This algorithm provided us with the regression slope predicted by the weighted Bayesian approach for each trial in each experimental condition of our experiments. Next, we calculated the response bias of such a model (Figure S3 in the online supplemental material) in the exact same way we did for our participants. For comparison, we also plotted in Figure S3 the bias shown by a classic Deming regression algorithm. Indeed, Deming regression is also thought to be more robust to outliers than ordinary least squares, because outlier data points affect (i.e., "pull") the regression line to a smaller extent

when they are orthogonally projected to it (as in Deming) than when they are vertically projected to it (as in OLS).

The results show that Deming regression, once again, nicely mimics participants' performance in Experiment 1, where participants were not explicitly told about outliers, but not Experiments 2 and 3. In other words, even if Deming regression is partially robust to outliers, its robustness is modest and both participants (Figure 6C) and Deming do not automatically exclude even distant outliers. However, we can see that the Bayesian model (set to probabilistically detect outliers beyond a threshold of $z = 2$) is much more robust to outliers and shows a behavior partially similar to humans in Experiments 2 and 3: for highest z-scores larger than 2.8, its bias stops increasing. Crucially, however, a difference remains: whereas in humans such bias ultimately decreases as the z-score becomes very large (Figure 6C), the bias for the model remains essentially flat for increasing values of z-scores. The results indicate that the Bayesian model, while close to humans, still differs from them in that it misses a mechanism to sharply reject obvious outliers.

Discussion

Across three experiments manipulating the number and distance of outliers in scatterplots and the level of attention toward them, we probed the human capacity for intuitive statistics in tasks of trend judgment, line fitting, and outlier detection, investigating whether outlier items are spontaneously included (as suggested by the literature on graph perception) or rather excluded from any statistical judgment (as predicted by the literature on ensemble perception). We now examine how the results provided answers to the five research questions presented in the introduction. We also try to integrate our findings, both with previous findings in the narrow domain of scatterplot perception and with the larger literature on ensemble and outlier perception (which did not use graphs as stimuli); indeed, as argued by Rensink (2021), studies on graphical representations can provide fruitful insights not just for graph perception but also, more broadly, for vision sciences.

First, do subjects spontaneously reject outliers when asked to perform a trend judgment or a regression estimation on a graph, without being told that there might be outliers? Experiment 1 is quite clear: participants do not spontaneously reject outliers and they integrate these deviant points in both their trend judgments and their regression estimations. As summarized in the introduction, recent studies on ensemble perception (e.g., Epstein et al., 2020; Haberman & Whitney, 2010) showed that, on the contrary, deviant items are easily discarded when participants are asked to provide an estimate of the average of a set. This contradiction might suggest that the intuitive extraction of visual statistics from a graph is not solely a form of ensemble perception. Indeed, when asked to fit a line or extract a trend from a graph, our participants performed a computation that goes beyond the simple "averaging" of a value on a common scale, as is the case for the ensemble perception of items of different hues or orientations. In these cases, the averaging is over a factor that is already present in each individual item: the average color of all items' color, the average orientation of all items' orientations (Whitney & Yamanashi Leib, 2018). In the case of scatterplots, the average item location is useless when assessing a trend, which arises from the relations between data points. Future research should try to disentangle the commonalities and differences between graph and ensemble perception (for a review: Cui & Liu, 2021). At the very least, our studies prove that the two processes are not fully overlapping. It is important to point out that, both in our stimuli and in the reviewed papers on ensemble perception, when multiple outliers were present, they were correlated with each other: more specifically, they either had the exact same level of deviation from the average value (Haberman & Whitney, 2010) or they were generated from a secondary value with the addition of random noise (Epstein et al., 2020), as was also the case here. Future research should investigate whether the same results hold (both for graph and ensemble perception) if the outliers are fully uncorrelated.

Second, do the number of outliers and their distance from the main dataset modulate the bias they induce? Our results from Experiments 1 and 2 show that yes, participants' errors and response times in the trend judgment task increase for a higher number of outliers and for a larger distance of these outliers from the main dataset. Likewise, the participants' slope estimates become increasingly biased (i.e., attracted toward outliers) for larger values of these factors. It is worth noting that those increases in error rate, response time, and response bias were significantly less pronounced for a main slope of $|.5|$ than for a shallower main slope of $|.25|$. In other words, when the main trend was steeper, outliers were less likely to affect participants' responses. This result makes sense: it is when the decision is most difficult, because the main slope is less pronounced, that outliers have the greatest influence. However, the effect of outliers on response times was still significant even when slope was regressed out, a finding that suggests a serial processing of outliers, with a cost of $\sim$20 ms per item. Overall, our findings extend previous research on outlier processing in scatterplots (Bobko & Karren, 1979; Correll & Heer, 2017; Meyer & Shinar, 1992; Meyer et al., 1997) by showing that deviant points in a scatterplot affect the human capacity for mental regression more if they are numerous and further from the main dataset.

One might argue that the stimuli we used comprised a too small number of observations (18), which may not be sufficient to allow the viewer to form a reliable mental regression from which to detect deviant points. However, the results from Figure 2 clearly show that our stimuli comprised enough evidence for subjects to accurately detect the regression slope. The reason we opted for 18 datapoints is double: first, we showed in previous research that humans are able to reliably compute mental regressions with as few as six datapoints, with a performance close to optimal for data sets like the ones we used in the current study (i.e., with 18 points generated from slopes steeper than .2; Ciccione & Dehaene, 2021); second, we wanted to avoid conditions in which outliers could too easily pop out, making the task trivial. Future studies could investigate the effects of the overall number of datapoints on outlier detection by parametrically varying this factor.

Third, can the outlier-induced bias be mitigated by drawing attention to them? In the fast trend judgment (first task of Experiments 1 and 2), devoting attention to outliers did not significantly improve participants' performance (Figure 2). This finding suggests that an extraction of the overall trend (including outliers) occurs fast and automatically—indeed, our hypothesis for outlier rejection suggests that it could be a necessary step prior to outlier detection and rejection. However, the comparison of the response bias in the line adjustment task from the three conditions of

attention deployment (Experiment 1: none; Experiment 2: medium; Experiment 3: high) revealed that, yes, outliers are more easily rejected when participants are aware of their presence and invited to discard them. It is worth clarifying that this finding does not imply that attention is needed for outlier processing itself: indeed, our findings from Experiment 1 (no attention) clearly show that deviant items affect trend judgments and slope estimations even more if participants are not aware of their presence. In agreement with this, several studies showed that attention is not necessary for the perceptual processing of visual items (Kouider & Dehaene, 2007), which can still attract spatial attention even when subliminally perceived (Astle et al., 2010; Robitaille & Jolicoeur, 2006) and clearly deviating from the other items (Hsieh et al., 2011). However, our results are congruent with the finding that attention can modulate even subliminal processing (Kiefer & Brendel, 2006; Naccache et al., 2002).

When attention was deployed toward outliers (but, crucially, no rejection was asked), one study found that deviant items in size or brightness were integrated in judgments of average size or brightness and biased participants' judgments toward the outlier value (de Fockert & Marchant, 2008). Our findings show that this strong attraction, exerted by both unattended and attended outliers, can be reduced if participants are explicitly asked to exclude them, but Experiment 3 suggests that it is hard to fully eliminate—even when a single outlier was present, and it was explicitly detected, it kept an influence on the participants' estimates of regression slopes. An interesting question for future studies is to what extent this strong attraction is resistant to training: in fact, a recent study showed that the estimation of correlation in a scatterplot improved significantly following long perceptual training sessions with feedback (Cui et al., 2018).

Fourth, how does outlier detection work? In the first task of our third experiment, we found that correct detection of outliers improved for larger distances from the main dataset, but also for more numerous outliers. The latter result might be due to at least two different reasons: a larger number of outliers may increase the probability for at least one of them to be seen; and/or it may make them globally more salient and recognizable (Kinchla, 1977). Future studies could try to disentangle these two hypotheses.

Interestingly, outlier detection exhibited considerably slower response times than trend judgments on the whole set (Figure 3B; Figure 6D for a direct comparison): this observation replicates previous evidence that visual judgments about the average value of the items in a set are faster than the detection of deviant observations present in those sets (Hochstein et al., 2018). This finding agrees with our model, according to which the extraction of the scatterplot trend is a necessary step prior to outlier detection, since the latter is based on their deviation from the main trend. Indeed, the paradox of outliers' detection (Epstein et al., 2020) is that an outlier is defined as deviating from a summary statistic computed on the entire set, meaning that it cannot be computed without also extracting such a summary reference value. Therefore, the higher response times observed for outlier detection might be the result of a trend judgment phase followed by outlier detection per se. It should, however, be noted that, perhaps as a consequence of those successive stages, those response times were highly variable and therefore any conclusion should be drawn with great caution.

We also formulated an explicit model of outlier detection, and tested it against many alternative models. The model hypothesizes that outliers are detected based on their elevated z-score; i.e., their large distance to the regression line, relative to the typical distance of other data points. Participants would compute a z-score for each data point, and evaluate whether the highest of these z-scores exceeds a threshold of about 2. This model was supported by both response times and error analyses. In response times, we found a distance effect, whereby outlier detection became increasingly faster for stimuli whose highest z-score increasingly deviated from 2. This is exactly what the model predicts: for stimuli comprising points with smaller z-scores, the absence of outliers is quickly detected, whereas for stimuli with outliers with higher z-scores, their presence is recognized increasingly fast. Likewise, we found that the percentage of "yes" responses was best modeled as a function of the highest z-score, with a sigmoidal function showing an inflection point around about 2. Importantly, the best fit was obtained when the z-score was calculated as the ratio between the orthogonal distance of the data point to the Deming fit, and the prescribed standard deviation of the datasets (i.e., the "noise" level). The explanatory advantage of the orthogonal distance over the vertical distance from OLS replicates our previous results showing that participants minimize the perpendicular euclidean distance of each point to the best-fitting line when computing a trend (Ciccione & Dehaene, 2021). On the other hand, the explanatory advantage of the prescribed standard deviation over the actual standard deviation of each stimulus merits a brief discussion. It might have been rational for participants to compute the actual noise level in every individual scatterplot in order to determine if a point is or not an outlier. However, humans are remarkably accurate at encoding the variability in a set of items (Morgan et al., 2008; Solomon, 2010) and they do so automatically, even when not explicitly asked for it (Khayat & Hochstein, 2018). Furthermore, the standard deviation of orthogonal distances from the fit seems to be used by humans when asked to perform correlation judgments (Yang et al., 2019). Therefore, it is reasonable to speculate that participants in our experiment computed the average noise level across trials—i.e., the prescribed standard deviation—and used it as their reference against which outliers were compared.

It is worth highlighting that we do not claim that humans are using explicit mental calculation to compute the z-score of each datapoint in the scatterplot. Indeed, the observed responses times would be incompatible with such a slow procedure. Our data simply suggest that, during fast graph perception tasks, humans deploy a fast process that tightly approximates a statistical model computing z-scores. As reviewed throughout the article, the human visual system is known to be able to compute complex summary statistics over briefly presented sets of items: the automatic computation of z-scores merely adds to this set of computational abilities. However, whether or not the z-score hypothesis holds should be more precisely studied. Future research could manipulate, for instance, the noise level in successive graphs and asks (a) whether the actual noise level (i.e., the denominator in the z-score formula) can be computed on a trial-by-trial basis; and (b) whether an approximate division of dot distance by this noise estimate actually occurs and what is its accuracy. A more parsimonious hypothesis is simply that the human visual system recycles its ability to detect objects' contours and principal axes and applies it to graphs, by extracting an estimation of the posterior distribution of all possible graph's contours (which would obviously depend on how noisy the graph is). Each datapoint would then be perceived either as

part of such distribution (and therefore included in the trend estimation) or out of it (thus detected as an outlier).

Fifth, finally, if outliers are correctly detected, does this mean that they can also be rejected? Experiment 3 concludes to the negative: outlier detection does not necessarily lead to outlier rejection. When we modeled participants' bias as a function of the highest *z*-score in the dataset (Figure 6B), we found that correct detection of the presence of outliers approached 90% for a highest *z*-score of 2.8. However, the response bias in the subsequent regression estimation (in which participants were asked to reject outliers; Figure 6C) showed that, although the bias was reduced in Experiment 3 (high attention) as compared with the two other experiments (none or medium attention), it was at its peak for a highest *z*-score of 2.8. It is only for stimuli with a highest *z*-score larger than 3.6 (i.e., with at least one extreme outlier) that the bias disappeared.

Interestingly, we also showed that an optimal Bayesian model that assigns a lower weight to outliers on the basis of their *z*-score (therefore, without fully rejecting them) behaves somewhat similarly to our participants, suggesting that human outlier detection and rejection may be a probabilistic computation. However, in this Bayesian model, the bias does not decrease sufficiently for large *z*-scores, whereas the human bias almost disappears then. This discrepancy may be due to the fact that the model uses the actual noise in the dataset, rather than an estimate of noise averaged over several trials (as used by humans). In fact, for larger highest *z*-scores, when more than one outlier is present, the *z*-score of those outliers necessarily decreases because high *z*-scores increase the overall noise level and, as a consequence, decrease their weight in the regression. On the contrary, humans seem able to calibrate their rejections on the basis of the noise of the main generative process, as already discussed in a previous section.

Taken together, these findings suggest that outlier rejection depends on two factors: the degree of attention toward them, and their deviation from the main dataset. Both factors seem to influence participants in placing a threshold past which they would be more likely to consider a data point as an outlier, beyond the normal noise in the dataset. In other words, the same data point could be seen either as the result of normal variability in the graph or as a significantly deviant observation, depending on task instructions. However, even when participants were maximally invited to pay attention to outliers and to detect and reject them before performing any regression estimation (Experiment 3), nonextreme outliers still biased their performance, even when they were correctly detected. This finding suggests that, to some extent, mental regression may be cognitively impenetrable (Pylyshyn, 1999; Stokes, 2013); correctly detecting outliers does not prevent them from influencing the participants' mental regression estimates. We can reasonably conclude that outliers in a graph are not treated as sets of items, thus confirming that graph perception does not operate identically to ensemble perception. We speculate that trend judgment and regression estimation are fast and largely automatic and that outliers, if present and detected, are rejected at a later time, with cognitive effort and following a probabilistic computation. In support of this hypothesis, a recent fMRI study on the neural bases of outlier processing for sets of colored objects (Cant & Xu, 2020) found that voluntarily discarding outliers led to activations that were not confined to early visual areas but involved fronto-parietal areas. Thus, two different types of processes (Kahneman, 2003) seem to be deployed during graph perception. Visual perception, including the automatic computation of the principal axes of an object or a graph, seems to interact with higher level cognition, including the deliberate rejection of outliers, with the second process not always able to counteract the information coming from the first (Pylyshyn, 1999).

Lastly, it is important to point out that our experimental tasks focused solely on the psychophysical aspects of graph perception, and did not include any specification of the names, characteristics and meaning of the x and y variables, as one would expect from "real" bivariate graphical representations. It seems likely that participants would have behaved differently if the stimuli were referring to actual data; indeed, outliers are usually either included or rejected from main analyses depending on several factors, including the statistical framework adopted by the scientist (frequentist or Bayesian), the experimental procedure of data acquisition, the type of variables, and their meaning. While our studies investigated the perceptual stages of outlier detection and rejection, future work should also consider using more ecologically valid stimuli in order to evaluate to what extent explicit knowledge of the data affects participants' biases and their probability to include or reject outliers.

Evidence-Based Suggestions to Improve Data Visualization of Outliers in Scatterplots

Based on the findings presented in this article, we conclude by proposing a few suggestions to improve outlier detection and rejection in data visualizations. Since these guidelines are speculative, although evidence-based, future research should empirically test their utility through appropriate behavioral studies.

1) Given that outliers are not spontaneously rejected, it could be helpful to explicitly identify all datapoints that exceed a predetermined *z*-score deviation from the overall linear regression. For instance, they could be put in a different color or, preferably, a smaller size or luminance. Such a manipulation of size and luminance was shown to be successful at modifying people estimations in a barycenter task (Hong et al., 2021).

2) Since human mental regressions tend to be performed on the whole dataset, even when outliers are correctly detected, scatterplots could include both the regression applied to all points and the regression after exclusion of the points that exceed a predetermined *z*-score. The direct comparison of a robust regression with a nonrobust one could help make the discrepancy between the two models more salient to the reader.

3) Since outlier detection is better than outlier rejection, interactive visualizations may help. A regression line would first be calculated over the entire dataset, and then the user would select potential outliers. The regression slope would then instantly adapt to exclude those points, which would allow for an interactive, online visualization of how outlier rejection changes the regression. In this manner, the defects of human intuition would be supplemented by human–machine interaction.

References

Akaike, H. (1998). *Information theory and an extension of the maximum likelihood principle*. Springer. https://doi.org/10.1007/978-1-4612-1694-0_15

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, *15*(3), 122–131. https://doi.org/10.1016/j.tics.2011.01.003

Anscombe, F. J. (1960). Rejection of outliers. *Technometrics*, *2*(2), 123–146. https://doi.org/10.1080/00401706.1960.10489888

Astle, D. E., Nobre, A. C., & Scerif, G. (2010). Subliminally presented and stored objects capture spatial attention. *The Journal of Neuroscience*, *30*(10), 3567–3571. https://doi.org/10.1523/JNEUROSCI.5701-09.2010

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

Cant, J. S., & Xu, Y. (2020). One bad apple spoils the whole bushel: The neural basis of outlier processing. *NeuroImage*, *211*, Article 116629. https://doi.org/10.1016/j.neuroimage.2020.116629

Chaloner, K., & Brant, R. (1988). A Bayesian approach to outlier detection and residual analysis. *Biometrika*, *75*(4), 651–659. https://doi.org/10.1093/biomet/75.4.651

Ciccione, L., & Dehaene, S. (2021). Can humans perform mental regression on a graph? Accuracy and bias in the perception of scatterplots. *Cognitive Psychology*, *128*, Article 101406. https://doi.org/10.1016/j.cogpsych.2021.101406

Ciccione, L., Sablé-Meyer, M., & Dehaene, S. (2022). Analyzing the misperception of exponential growth in graphs. *Cognition*, *225*, Article 105112. https://doi.org/10.1016/j.cognition.2022.105112

Correll, M., & Heer, J. (2017). Regression by eye: Estimating trends in bivariate visualizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems–CHI '17* (pp. 1387–1396). Association for Computing Machinery. https://doi.org/10.1145/3025453.3025922

Cui, L., & Liu, Z. (2021). Synergy between research on ensemble perception, data visualization, and statistics education: A tutorial review. *Attention, Perception, & Psychophysics*, *83*(3), 1290–1311. https://doi.org/10.3758/s13414-020-02212-x

Cui, L., Massey, C. M., & Kellman, P. J. (2018, July 25–28). *Perceptual learning in correlation estimation: The role of learning category organization.* The 40th Annual Conference of the Cognitive Science Society (pp. 262–267). https://par.nsf.gov/servlets/purl/10254053

de Fockert, J. W., & Marchant, A. P. (2008). Attention modulates set representation by statistical properties. *Perception & Psychophysics*, *70*(5), 789–794. https://doi.org/10.3758/pp.70.5.789

Epstein, M. L., Quilty-Dunn, J., Mandelbaum, E., & Emmanouil, T. A. (2020). The outlier paradox: The role of iterative ensemble coding in discounting outliers. *Journal of Experimental Psychology: Human Perception and Performance*, *46*(11), 1267–1279. https://doi.org/10.1037/xhp0000857

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Friendly, M., & Denis, D. (2005). The early origins and development of the scatterplot. *Journal of the History of the Behavioral Sciences*, *41*(2), 103–130. https://doi.org/10.1002/jhbs.20078

Godau, C., Vogelgesang, T., & Gaschler, R. (2016). Perception of bar graphs—A biased impression? *Computers in Human Behavior*, *59*, 67–73. https://doi.org/10.1016/j.chb.2016.01.036

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. https://doi.org/10.1016/S0896-6273(02)00971-6

Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention, Perception, & Psychophysics*, *72*(7), 1825–1838. https://doi.org/10.3758/APP.72.7.1825

Hawkins, D. M. (1980). *Identification of outliers*. Chapman and Hall. https://doi.org/10.1007/978-94-015-3994-4

Hochstein, S., Pavlovskaya, M., Bonneh, Y. S., & Soroker, N. (2018). Comparing set summary statistics and outlier pop out in vision. *Journal of Vision*, *18*(13), Article 12. https://doi.org/10.1167/18.13.12

Hong, M.-H., Witt, J. K., & Szafir, D. A. (2021). *The weighted average illusion: Biases in perceived mean position in scatterplots*. ArXiv. http://arxiv.org/abs/2108.03766

Hsieh, P.-J., Colas, J. T., & Kanwisher, N. (2011). Pop-out without awareness: Unseen feature singletons capture attention only when top-down attention is available. *Psychological Science*, *22*(9), 1220–1226. https://doi.org/10.1177/0956797611419302

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *The American Economic Review*, *93*(5), 1449–1475. https://doi.org/10.1257/000282803322655392

Khayat, N., & Hochstein, S. (2018). Perceiving set mean and range: Automaticity and precision. *Journal of Vision*, *18*(9), Article 23. https://doi.org/10.1167/18.9.23

Kiefer, M., & Brendel, D. (2006). Attentional modulation of unconscious "automatic" processes: Evidence from event-related potentials in a masked priming paradigm. *Journal of Cognitive Neuroscience*, *18*(2), 184–198. https://doi.org/10.1162/089892906775783688

Kinchla, R. A. (1977). The role of structural redundancy in the perception of visual targets. *Perception & Psychophysics*, *22*(1), 19–30. https://doi.org/10.3758/BF03206076

Kouider, S., & Dehaene, S. (2007). Levels of processing during nonconscious perception: A critical review of visual masking. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, *362*(1481), 857–875. https://doi.org/10.1098/rstb.2007.2093

Liu, T., Li, X., Bao, C., Correll, M., Tu, C., Deussen, O., & Wang, Y. (2021). *Data-driven mark orientation for trend estimation in scatterplots.* Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–16. https://doi.org/10.1145/3411764.3445751

Meyer, J., & Shinar, D. (1992). Estimating correlations from scatterplots. *Human Factors*, *34*(3), 335–349. https://doi.org/10.1177/001872089203400307

Meyer, J., Taieb, M., & Flascher, I. (1997). Correlation estimates as perceptual judgments. *Journal of Experimental Psychology: Applied*, *3*(1), 3–20. https://doi.org/10.1037/1076-898X.3.1.3

Micallef, L., Palmas, G., Oulasvirta, A., & Weinkauf, T. (2017). Towards perceptual optimization of the visual design of scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, *23*(6), 1588–1599. https://doi.org/10.1109/TVCG.2017.2674978

Morgan, M., Chubb, C., & Solomon, J. A. (2008). A 'dipper' function for texture discrimination based on orientation variance. *Journal of Vision*, *8*(11), 9. https://doi.org/10.1167/8.11.9

Naccache, L., Blandin, E., & Dehaene, S. (2002). Unconscious masked priming depends on temporal attention. *Psychological Science*, *13*, 416–424. https://doi.org/10.1111/1467-9280.0047

Orr, J. M., Sackett, P. R., & Dubois, C. L. Z. (1991). Outlier detection and treatment in I/O psychology: A survey of researcher beliefs and an empirical illustration. *Personnel Psychology*, *44*(3), 473–486. https://doi.org/10.1111/j.1744-6570.1991.tb02401.x

Pastore, M., Lionetti, F., & Altoè, G. (2017). When one shape does not fit all: A commentary essay on the use of graphs in psychological research. *Frontiers in Psychology*, *8*, Article 1666. https://doi.org/10.3389/fpsyg.2017.01666

Pylyshyn, Z. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences*, *22*(3), 341–365. https://doi.org/10.1017/S0140525X99002022

Reimann, D., Blech, C., & Gaschler, R. (2020). Visual model fit estimation in scatterplots and distribution of attention. *Experimental Psychology*, *67*(5), 292–302. https://doi.org/10.1027/1618-3169/a000499

Rensink, R. A. (2021). Visualization as a stimulus domain for vision science. *Journal of Vision*, *21*(8), Article 3. https://doi.org/10.1167/jov.21.8.3

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

Robitaille, N., & Jolicoeur, P. (2006). Fundamental properties of the N2pc as an index of spatial attention: Effects of masking. *Canadian Journal of Experimental Psychology*, *60*(2), 101–111. https://doi.org/10.1037/cjep2006011

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

Smiti, A. (2020). A critical overview of outlier detection methods. *Computer Science Review*, *38*, Article 100306. https://doi.org/10.1016/j.cosrev.2020.100306

Solomon, J. A. (2010). Visual discrimination of orientation statistics in crowded and uncrowded arrays. *Journal of Vision*, *10*(14), Article 19. https://doi.org/10.1167/10.14.19

Stokes, D. (2013). Cognitive penetrability of perception. *Philosophy Compass*, *8*(7), 646–663. https://doi.org/10.1111/phc3.12043

Sunday, M. A., Patel, P. A., Dodd, M. D., & Gauthier, I. (2019). Gender and hometown population density interact to predict face recognition ability. *Vision Research*, *163*, 14–23. https://doi.org/10.1016/j.visres.2019.08.006

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble perception. *Annual Review of Psychology*, *69*(1), 105–129. https://doi.org/10.1146/annurev-psych-010416-044232

Yang, F., Harrison, L. T., Rensink, R. A., Franconeri, S. L., & Chang, R. (2019). Correlation judgment and visualization features: A comparative study. *IEEE Transactions on Visualization and Computer Graphics*, *25*(3), 1474–1488. https://doi.org/10.1109/TVCG.2018.2810918

Received April 17, 2022
Revision received July 19, 2022
Accepted August 24, 2022 ■