Research Article

Visual Model Fit Estimation in Scatterplots and Distribution of Attention

Influence of Slope and Noise Level

Daniel Reimann, Christine Blech, and Robert Gaschler

Department of Psychology, FernUniversität in Hagen, Hagen, Germany

Abstract. Scatterplots are ubiquitous data graphs and can be used to depict how well data fit to a quantitative theory. We investigated which information is used for such estimates. In Experiment 1 (N = 25), we tested the influence of slope and noise on perceived fit between a linear model and data points. Additionally, eye tracking was used to analyze the deployment of attention. Visual fit estimation might mimic one or the other statistical estimate: If participants were influenced by noise only, this would suggest that their subjective judgment was similar to root mean square error. If slope was relevant, subjective estimation would mimic variance explained. While the influence of noise on estimated fit was stronger, we also found an influence of slope. As most of the fixations fell into the center of the scatterplot, in Experiment 2 (N = 51), we tested whether location of noise affects judgment. Indeed, high noise influenced the judgment of fit more strongly if it was located in the middle of the scatterplot. Visual fit estimates seem to be driven by the center of the scatterplot and to mimic variance explained.

Keywords: fit estimations, perception, data graphs, scatterplots

In many professional activities, such as quantitative research and engineering, people need to determine how well data points fit to a theoretical prediction. For instance, in science studies (e.g., Brewer, 2012), it has been pointed out that researchers hardly observe their object of interest directly and rarely check their theoretical predictions against direct observations. Instead, they make observations by using data graphs identifying relevant characteristics of the object of interest and for testing predictions. Bogen and Woodward (1992) suggested that it is data rather than perceptual beliefs that play a central evidential role to current science. Thus, while the early study of perception and psychophysics was driven by the challenges early astronomers faced (e.g., Brewer, 2012), turning away from direct observation and engaging with processed data instead might suggest that perceptual limitations no longer limit science. However, Brewer (2012) and others (cf. Wickham et al., 2015) argued that we now need to focus on better understanding how evidence is perceived against the background of theories in data graphs. Using graphs — as opposed to tables with numbers — allows us to harvest the computational power of the visual system to apprehend relations with little effort (e.g., Schnotz & Bannert, 2003). In some cases, however, the estimates the visual system provides are systematically biased (cf. Godau et al., 2016).

Given the substantial use of data graphs in conducting and communicating research (cf. Smith et al., 2000, 2002), we need to better understand how people use data graphs to weigh scientific evidence and theories. The studies of Smith et al. suggest that there is preference for visual presentation of data in harder sciences. Particularly in the natural sciences, readers of articles might, by a large share, use data graphs to judge the fit between theory and data. Such a visual judgment of fit (in addition to indices) is warranted. For instance, in the literature on skill acquisition, different variants of chunking-based learning could be pinned down to the prediction of a power law versus a negatively accelerated exponential learning curve (e.g., Evans et al., 2018; Heathcote et al., 2000). Exclusively considering fit indices could cause the observer to overlook important information that would otherwise be apparent in the graph: One theory might systematically overestimate or underestimate the asymptote (cf. Palmeri, 1999).

Visual model descriptions are recommended in the statistical literature to enable subjective fit estimations as adjuncts to numerical summaries (e.g., Wickham et al., 2015). Many studies (e.g., Evans et al., 2018) use scatterplots to visually present a quantitative theory and its associated data. Scatterplots have been described as the most useful invention in the history of data graphs and have found their way into the public sphere through public media (Bergstrom & West, 2018; Friendly & Denis, 2005).

While there is a long tradition of research on fit indices quantifying how well data fit a theoretical prediction (e.g., Pitt et al., 2002; Roberts & Pashler, 2000; Wagenmakers, 2003), we know little on how fit is estimated by viewers when prediction and data are displayed visually in scatterplots. In case of a simple linear model, first insights might be derived from the vast literature on how individuals estimate correlations from scatterplots.

Some studies suggested that people use the shortest (perpendicular) distance between data point and the regression line (90° angle) rather than the vertical distance (parallel to the *y*-axis) for estimations of correlations (e.g., Meyer et al., 1997; Yang et al., 2019). This visual approach is in contrast to most statistical procedures that usually rely on the vertical distances. A range of studies (cf. Doherty & Anderson, 2009) identified properties of a scatterplot which are unrelated to the statistical correlation but influence the perceived strength of the relationship. Accordingly, factors influencing the judgments include properties of the axis (scaling and theory-relevance of the labels), the point cloud (density, shape, size and number of the points, and the presence of outliers), and the regression line (mere presence and slope). Some of these aspects can be manipulated by the many design choices in scatterplots (cf. Sarikaya & Gleicher, 2018).

Lane et al. (1985) investigated how different combinations of correlation-related components (error variance, slope, and variance of *x*) affect judgments of correlation. The authors found higher estimates of correlation for scatterplots with lower error variance, higher variance in *x*, and steeper slopes. A comparison of the different influences revealed that the error variance had the strongest influence.

The influence of slope was also investigated in an experiment by Meyer and Shinar (1992). Participants estimated the strength of association in scatterplots with different slopes of the regression lines (slopes 30°, 45°, and 60°). Higher estimates of correlation resulted for plots with a *shallower* slope. Similarly, two experiments by Meyer et al. (1997) estimated the strength of the correlation in scatterplots with slopes ranging from 22° to 55°, revealed higher ratings for shallower slopes. Across the studies, Meyer et al. explained this effect as a side effect of their approach in manipulating the slope. In order to change the slope of the regression line, they changed the scales of the axes, which led to a lower density of the point cloud for steeper slopes. Consequently, in scatterplots with a steeper slope, the vertical distances of the data points to the line increased.

Another feature that can affect the perception of correlation is the mere presence of the regression line. Several studies (Meyer et al. 1997; Meyer & Shinar, 1992) demonstrated that its presence can lead to higher estimations of association. The regression line might serve as a perceptual center that increases perceived correlation.

Many studies (e.g., Bobko & Karren, 1979; Cleveland et al., 1982; Lauer & Post, 1989; Rensink, 2017; Rensink & Baldridge, 2010; Strahan & Hansen, 1978) showed that viewers tend to underestimate the strength of the correlation. As this bias differs depending on correlation strength, researchers targeted the psychophysical function. Some studies (e.g., Bobko & Karren, 1979; Rensink & Baldridge, 2010) found a positively accelerated shape between the presented correlation and the perceived correlation. Perception is more sensitive to changes in higher correlations than to changes in lower correlations. The usual task in these studies requires the viewer to determine the correlation coefficient for scatterplots with different degrees of correlation strengths. The tendency to underestimate has been replicated with different methodological approaches. For example, in some studies (e.g., Bobko & Karren, 1979), participants were asked to give direct numerical estimates. In more recent studies with similar findings (e.g., Rensink & Baldridge, 2010), participants had to adjust the correlation with a slider so that it was exactly halfway between two reference scatterplots. Taking into account the consistent finding of (a) the underestimation of the correlation and (b) that showing the regression line increases the estimate has led to the frequent recommendation to add the line as a default setting in scatterplots (e.g., Doherty & Anderson, 2009).

The literature on visual estimation of correlation provides valuable information about the perception of data in regard to an important numerical measure of goodness of fit (*r*). Yet, this evidence is only indirect as work on the estimation of correlation in scatterplots deals with the relationship between the *x*- and *y*-coordinates of the data points rather than with the relationship of data points and a visually presented model line. Given that a regression line seems to influence correlation estimates even when the task is to estimate the correlation among the data points (e.g., Doherty & Anderson, 2009), rather than reporting on the relation between points and line, studies directly addressing visual estimation of fit between model line and data points seem warranted. In many practical situations of perceiving scatterplots with model lines, viewers are not instructed to view the graphs in relation to a particular fit coefficient, thus allowing for a rather intuitive grasp of fit. This is also relevant for laypeople who may not be familiar

with statistical coefficients. For example, instead of estimating the variance explained by the model relative to overall variance, one might only be interested in how much the data points deviate from the model line (noise). This would be in line with the root mean square error (RMSE; e.g., Schunn & Wallach, 2005). In fact, in many cases, for example, within the field of predictive modeling (Wickham et al., 2015), one central question is, how accurate the predictions of a model are? The relevance of such a task has even been pointed out in the literature on correlation estimation. For example, Lane et al. (1985) stated that in many natural situations, the most important feature with regard to any covariation may be the accuracy of the predicted *y*-value.

Thus, while previous work on the perception of scatterplots was primarily about the visual estimation of the statistical coefficient *r*, often without showing any model, our focus was on the subjective impression of fit between model and data. The aim of the present study was to investigate if and how strongly slope and noise affect subjective fit estimations between a linear model line and data. In order to hold the vertical distances (noise) constant across different slopes, we first constructed the slopes and then created new data points for each slope. Noise was constant across scatterplots of different slopes, but the total variance in *y* increased with steeper slopes. The noise (and thus RMSE) was manipulated by adding different values of vertical distances (see Method). Overall, the approach allowed to analyze the influence of slope and noise for models that make predictions for the same range of *x*-values.

We expected higher estimates of fit for scatterplots with lower noise and steeper slopes. The latter expectation is based on the idea that the perception could be dominated by the shortest distances to the line (as with correlation estimation; cf. Meyer et al., 1997) which are shorter for scatterplots with a steeper slope. The constructed graphs had a regression line of least squares. To clarify that the task was to determine the goodness of fit instead of evaluating how well the line reflects the line of least squares, we pointed out that a perfect fit would be to have all data points on the line. This allowed us to test whether participants' judgments would mimic RMSE or would (in addition) resemble variance explained.

An additional aim of the study was to investigate the distribution of attention on subjective fit estimations and its possible consequences for the ratings. We decided to use eye tracking as an objective measure of visual attention localization as a prerequisite for understanding the underlying mechanisms of information processing. Using this approach made it also possible to examine whether there is any kind of perceptual center of gravity. The idea of a perceptual center in scatterplots has been considered in correlation estimation research (e.g., Meyer & Shinar, 1992).

Experiment 1

Method

Participants

Twenty-five German-speaking psychology students (18 women, 7 men, age $M = 32.9$ years, $SD = 10.8$) participated in the experiment. The number was based on a power analysis with a medium effect size of .30 for *f*, an α of .05, and a power of .90 for within-subjects ANOVA main effects (3 measurements) with G*power (Faul et al., 2009). Participants received course credit for compensation.

Materials and Procedure

Participants were tested individually in the laboratory of the FernUniversität in Hagen. The procedure was approved by the ethics review board of the faculty of psychology at the FernUniversität. After obtaining informed consent, participants were seated in front of the eye tracker. They were instructed to determine how well the data points fitted to the shown line by using the mouse to click on an analogue scale (0–100), which was shown below each scatterplot. Its endpoints had the labels "very bad" on the left and "very good" on the right. The axes of the scatterplots were unlabeled. After the perfect fit explanation, each person saw 36 scatterplots in a randomized order on a computer screen. The scatterplots were constructed based on a 3 (slope) × 3 (noise) × 4 (random pattern) within-subjects design.

There were three different slopes (22.5°, 45°, and 67.5°) and three levels of noise ($SD = 1$, 2, and 3; see below). For each combination of slope and noise (shown in Figure 1), we generated four parallel versions of scatterplots. We first created the line and then the data points in each scatterplot. We determined the slope with the slope coefficients of the model lines: A slope of 0.5 was equivalent to an angle of 22.5°, a slope of 1 was equivalent to an angle of 45°, and a slope of two was equivalent to an angle of 67.5°. For each of seven points along the *x*-axis, three values for the *y*-axis were calculated independently. The three *y*-values were drawn randomly and then standardized (subtracting the mean and dividing by the *SD*). This led to a mean of zero and a *SD* of one. In order to change the *SD* to the desired noise level, we simply multiplied the three values with a number (1, 2, or 3). Hence, the resulting *SD* of the three *y*-values was 1, 2, or 3 for scatterplots with small, medium, and large amounts of noise, respectively. In order to place the mean of the three *y*-values exactly on the line,

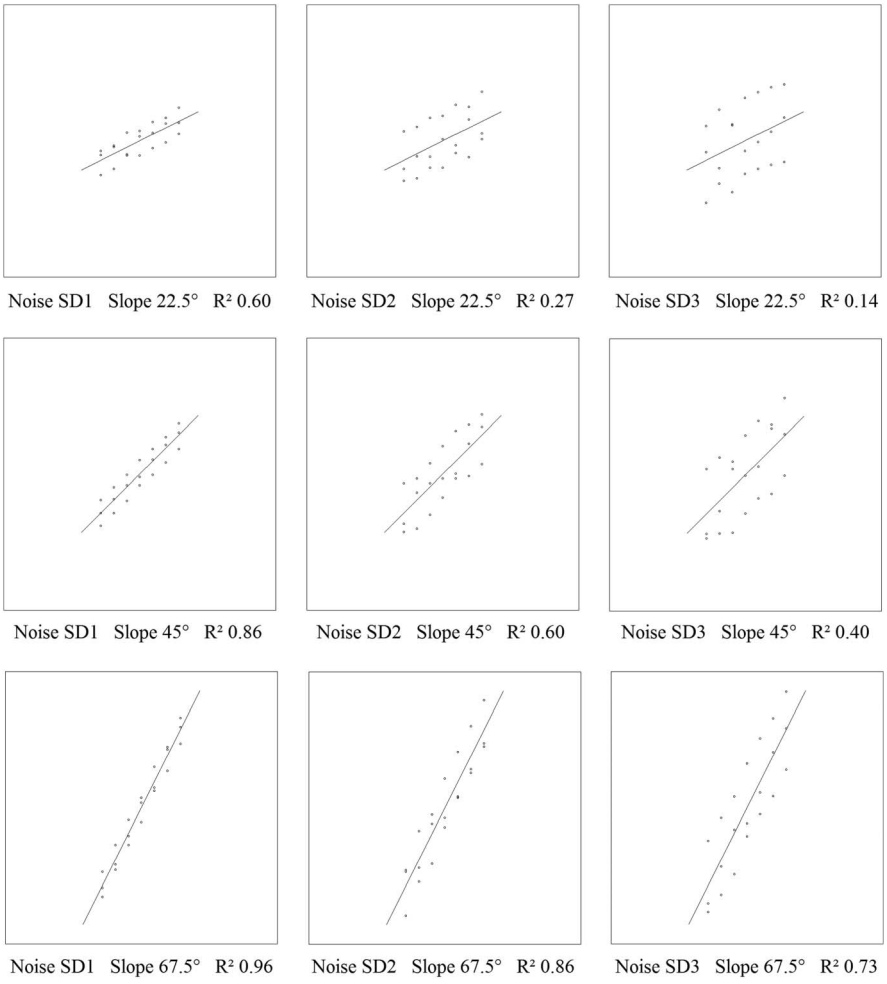

Figure 1. Example scatterplots for each combination of slope and noise. Each model makes predictions for the same range of *x*-values.

we then added a constant value to the three values (depending on the *x*-value). Assigning three *y*-values for each of the seven *x*-positions led to 21 data points for each scatterplot. Due to this construction approach, all resulting scatterplots had exactly the mentioned slopes and *SD*s. The script for the construction of stimuli, the stimuli, data and scripts of the present study are available online (Reimann, 2019).

The scatterplots were shown in the program ExperimentBuilder by SensoMotoric Instruments (SMI). A screen-based eye tracker (SMI RED 250 Hz) was used. After being informed about the purpose of the study, participants began the experiment. They were seated at 60 cm distance from the 24-inch monitor. A nine-point calibration was conducted. In each trial, a diagram and a scale for fit estimation were presented and fixations were recorded while the participant viewed the scatterplot. When the participant had indicated subjective fit by a mouse click on the scale, the experimenter switched on the next stimulus by pressing the space key.

The dependent variables were (1) rating of the subjective fit between the line and the data points and (2) the percentage of fixations in the middle versus at the borders of the scatterplot.

Results

Ratings

Figure 2 shows the mean values for the ratings of each scatterplot configuration ranging from 27.35 (slope 22.5°, noise 3 *SD*) to 88.11 (slope 67.5°, noise 1 *SD*). As we presented four different stimuli for each combination of noise level and slope, we could estimate reliability. It was above .71 for all combinations (see Appendix Table A1 for details). As to be expected, Figure 2 suggests that higher noise led to lower fit ratings. Importantly, the fit was rated higher for scatterplots with a steeper slope.

A 3 × 3 ANOVA with the within-subject factors slope (22.5°, 45°, and 67.5°) and noise (*SD* = 1, *SD* = 2, and *SD* = 3)

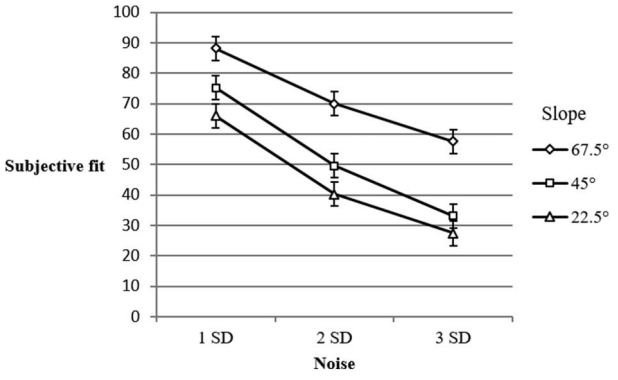

Figure 2. Mean values in rating for each configuration of a scatterplot. Error bars indicate the 95% within subjects CI according to Masson and Loftus (2003) based on the pooled error term.

showed a significant main effect for slope, F (1.32, 31.57) = 112.27, $p < .001$, $\eta_p^2 = .82$, a significant main effect for noise, F(1.19, 28.54) = 154.62, $p < .001$, $\eta_p^2 = .87$ (both Greenhouse–Geisser corrected), and an interaction effect, F(4, 96) = 5.01, $p = .001$, $\eta_p^2 = .17$. For the main effects, all pairwise comparisons (Bonferroni-corrected) were significant, $p < .001$. The interaction effect showed a lower influence of noise for the steepest slope. The difference between the mean values of noise 1 *SD* and noise 3 *SD* was 30.6 for the for a slope of 67.5°, 42.13 for a slope of 45°, and 38.61 for a slope of 22.5°.

Noise and slope did not only have effects in the average of the sample. An analysis on the individual level showed that each participant rated the steepest slope higher than the shallowest slope and the lowest noise higher than the highest noise. The mean difference in estimations between the highest noise and the lowest noise ($M = 37.11$, $SD = 14.25$) was significantly higher than the difference between the steepest slope and the shallowest slope ($M = 27.37$, $SD = 11.84$), $t(24) = -4.04$, $p = <.001$, $d = 0.73$. Thus, the impact of the noise manipulation on the fit rating was larger than the impact of the experimental variation of slope.

The difference in the size of the effect would be especially informative if it could be related to a common metric of the extent to which the independent variable was manipulated. In order to directly compare the influence of noise and slope on the fit estimation, we quantified the relative influence with the aid of a common scale (explained variance by the linear model). For a slope of 22.5°, the mean explained variance was 33.85% (mean across three noise levels). For a slope of 67.5°, the mean explained variance was 84.81%. Taking the difference (84.81% – 33.85% = 50.96%) suggests that the slope manipulation spanned a range of 50.96% of explained variance. The mean difference in fit rating between the steepest slope and the shallowest slope was 27.37. The quotient (27.37 by 50.96%) suggests that 0.54 rating points were gained per percent of explained variance in the scatterplot. In order to compare this measure of impact to the influence found for the variation of noise, we calculated the mean explained variance for *SD* 1 (mean across three slopes = 80.57%) and for *SD* 3 ($M = 42.34\%$). The difference in explained variance between the high and low noise scatterplots was 38.23%. Dividing the corresponding difference in the fit rating of ($M = 37.11$) this value resulted in an impact of 0.97 rating points per percent of explained variance in the scatterplots. Thus, when comparing the strength of the two independent variables on a common scale, again, the impact of noise was larger than the impact of slope.

Fixations

Figure 3 shows the fixations across all 25 participants for each scatterplot type. Visual inspection suggested that the fixations tended to fall on the center of the scatterplot and that this pattern seemed to be stronger for scatterplots with a shallower slope.

In order to be able to compare the degree of clustering of the fixations for different levels of slope and noise, we split each scatterplot into three equal areas (Figure 4) along the *x*-axis from the beginning to the end of the regression line and calculated the percentage of fixations for each area, excluding the remaining space on the left and right margins. As participants differed in the number of fixations, before determining the average over all subjects, we first calculated percentages for each participant individually so that each single subject had the same weight in the group-average results. Table 1 shows the outcome. The 3 × 3 ANOVA for Area 2 percentage as a dependent variable with the within-subject factors slope (22.5°, 45°, and 67.5°) and noise (*SD* = 1, *SD* = 2, and *SD* = 3) showed a significant main effect for slope, $F(2, 48) = 33.07$, $p < .001$, $\eta_p^2 = .58$. There was neither a main effect for noise, $F(2, 48) = 0.68$, $p = .51$, $\eta_p^2 = .03$, nor a significant interaction effect, $F(2.58, 61.8) = 2.13$, $p = .11$, $\eta_p^2 = .08$. While the average percentage of fixations falling in the center was high for scatterplots with a shallow slope of 22.5° ($M = 88\%$) and with a slope of 45° ($M = 86\%$), clustering in the middle was reduced to $M = 75\%$ for scatterplots with a steep slope of 67.5°.

Discussion

As expected, the subjective fit estimations were lower for scatterplots with higher noise. The result that steeper slopes led to higher fit estimations suggests that participants not only took the vertical distances into account (as in the numerical fit measure RMSE). It is therefore possible that the perception of fit was dominated by the perpendicular distances from the data points to the line. This

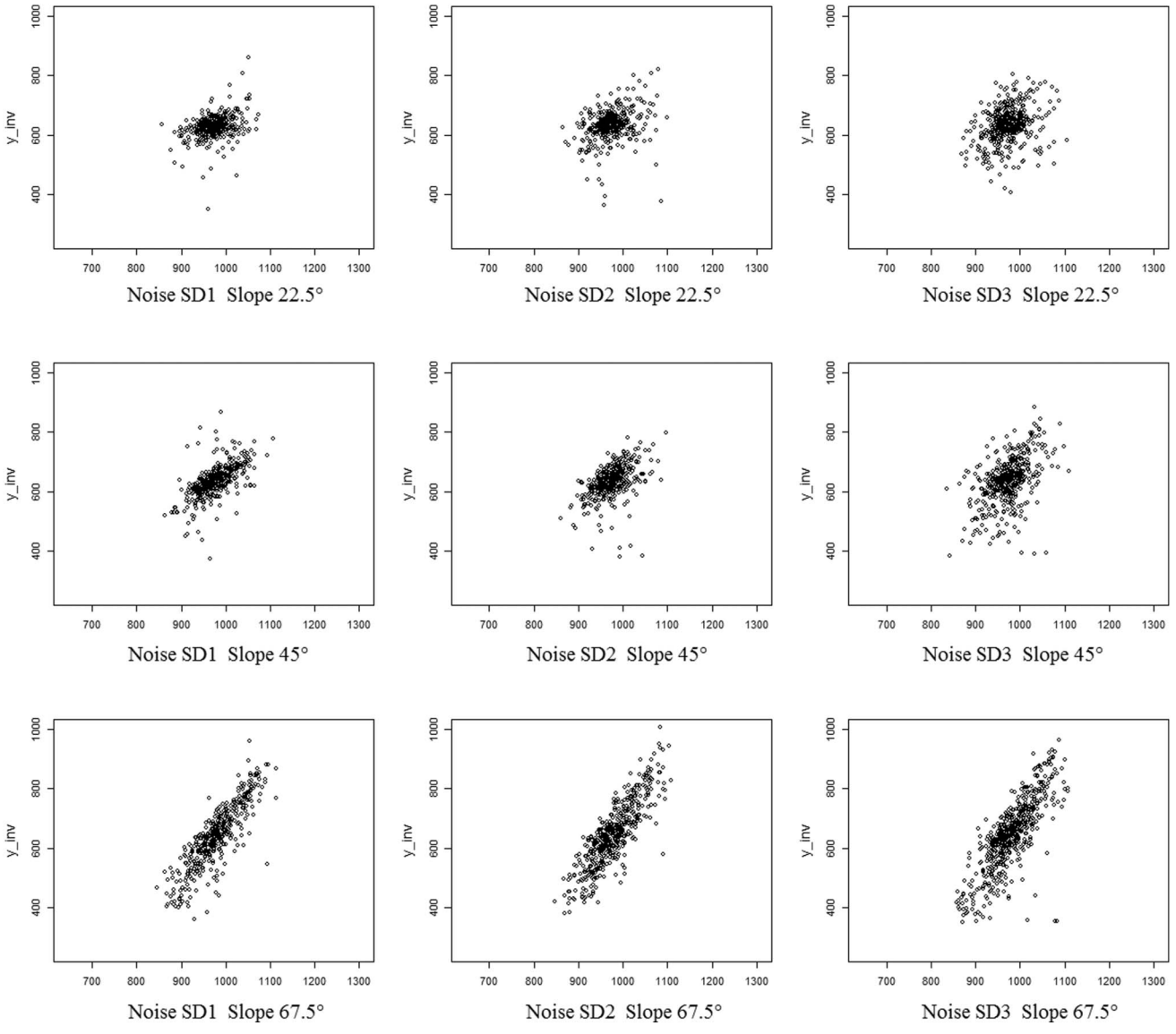

Figure 3. Fixations for each scatterplot type across all 25 participants. Numbers along the axes refer to the screen coordinates (1,920 × 1,080). The perceived scatterplots were in the center of the screen and did neither have numbers nor information about the level of noise and slope. For the alignment between stimuli and screen, see file screen_stimuli_alignment in Reimann (2019). Furthermore, we provide a heatmap for each of the 36 stimuli online.

distance has also been contemplated as a proxy that viewers use to estimate correlation (e.g., Meyer at al., 1997; Yang et al., 2019). The effect sizes of noise (η_p^2 = .87) and slope (η_p^2 = .82) were similarly high. However, allocating the manipulation of both variables on a common scale of explained variance suggested that the relative influence of noise was approximately twice as strong as the influence of the slope. The influence of noise and higher estimates for graphs with steeper slopes are in line with the correlation estimates reported in Lane et al. (1985). Only at first sight the slope effect seems inconsistent with the findings by Meyer and Shinar (1992) and Meyer et al. (1997), where steeper slopes led to lower estimations of correlation. While our manipulation of slope did not change the vertical distances between data points and prediction line, Meyer et al.'s approach of changing the axes led to a lower density of the point cloud and reduced vertical distances for shallower slopes. In contrast to the latter, our approach allowed us to analyze the influence of slope while keeping noise constant. Similar efforts have been made by Lane et al. (1985), but the authors did not show a model and focused explicitly on correlation.

The results of the fixations revealed that the majority of fixations fell into the center (Area 2) and that there was a tendency to also focus on the outer areas in the cases of steeper scatterplots. Stronger consideration of the outer areas

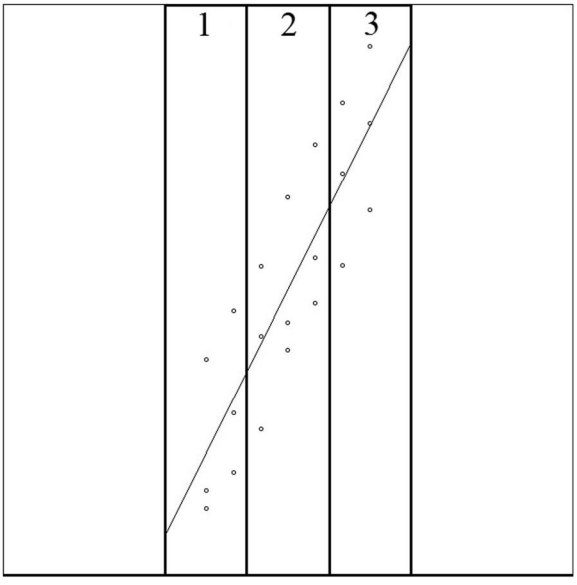

Figure 4. Splitting the scatterplots into three areas.

for steeper slopes might have occurred due to its longer regression lines (a consequence of the endeavor to compare models with predictions for the same range of *x*-values).

The tendency to fixate on the center of an image (center bias) is a well-known effect in the scene viewing literature (e.g., Tatler, 2007) and is explained as an optimal viewing position for effective exploring. Tatler showed that the bias is very robust and does not disappear when the salient image features are placed outside the center.

Experiment 1 only included scatterplots with a fairly homogenous point pattern along the line. In some applications, models can have area-specific deviations of fit (e.g., Wickham et al., 2015). In order to investigate a possible consequence of clustering, we decided to conduct a second experiment with different patterns.

Experiment 2

In Experiment 1, we observed that participants mostly fixated on the center of the scatterplots (Area 2). As the level of noise was equal in and outside the center allocating attention to the center, a further experiment was needed in order to test whether this allocation of attention was consequential for the fit ratings. If the data points in the center have more influence on the rating than the points in the periphery, then higher noise would have a larger impact on ratings if they are located in the center rather than outside the center of the scatterplot. We turned the analysis approach from Experiment 1 into an experimental manipulation, dividing each scatterplot into three areas as defined in Experiment 1 and implementing different noise levels for different areas within one scatterplot. We expected that the estimations would be influenced more by the point pattern of the center as opposed to the periphery.

Visual inspection of plots with such deviations is relevant when checking for the assumption of homoscedasticity in regressions, which means that the variance of errors is the same across all levels of the independent variable (Osborne & Waters, 2002).

Table 1. Distribution of fixations for each area in percentage

Slope	Area 1 *M* (*SD*)	Area 2 *M* (*SD*)	Area 3 *M* (*SD*)
22.5°	6.37 (6.62)	87.93 (8.92)	5.69 (7.68)
45°	6.78 (5.02)	85.82 (8.90)	7.40 (7.27)
67.5°	10.35 (8.16)	74.82 (12.89)	14.81 (12.91)

Means and *SD*s.

Method

Participants

Fifty-one German-speaking psychology students (32 women, 19 men, age $M = 34.6$ years, $SD = 9.8$) participated for course credit in the experiment.

Materials and Procedure

The experiment was conducted as an online experiment with integrated visual stimuli, programmed with the tool Unipark. Each person saw 36 scatterplots in a randomized order. The variation between the scatterplots followed a 3 (area with highest noise) × 2 (overall noise) within-subjects design with six scatterplots per design cell. The diagram axes were scaled from 0 to 10 on 2-point intervals, and the slope remained the same across all scatterplots. Ticks were added to the axes to account for a more realistic scenario. For the slope, we chose the value with the highest percentage of fixations in the center in Experiment 1 (22.5°) to maximize the chances for obtaining a strong effect. We used a 10 × 10 coordinate system and a linear equation of 2.5 + 0.5x. The instructions and construction approach for the scatterplots followed the logic of Experiment 1 (except for the fact that there were nine instead of seven points on the *x*-axis). We manipulated the location of area with highest noise, which was four times as high as the noise for the remaining two areas. For example, when two areas of the scatterplot had a noise of 0.25, the area with the highest noise had a noise of 1 (see examples in the upper line in Figure 5). The location of the area with the highest noise could be left, middle, or right. For this, we used high

noise for the first three, middle three, or last three values on the *x*-axis, respectively.

Additionally, we manipulated the overall noise without changing the ratio (one to four) between areas of lower and high noise within one scatterplot. The values for the *SD* for the lower overall noise were 1 (Area 1), 0.25 (Area 2), and 0.25 (Area 3) and for the high overall noise 2 (Area 1), 0.5 (Area 2), and 0.5 (Area 3) as examples of scatterplots where the area with highest deviations was on the left side. For each combination of overall noise (high vs. low) and area with highest noise (Area 1, Area 2, and Area 3), we used six different randomly generated scatterplots, resulting in 36 stimuli (see Figure 5 for examples). Stimuli are available online (Reimann, 2019).

Results

Figure 6 shows the mean values for the ratings of each scatterplot configuration. All scales had a reliability above .92 (see Appendix Table A2 for details). A 2 × 3 ANOVA with the within-subject factors overall noise (high and low) and area with the highest noise (Area 1, Area 2, and Area 3) showed a significant main effect for overall noise, $F(1, 50) = 155.18$, $p < .001$, $\eta_p^2 = .76$, a significant main effect for area with the highest noise, $F(2, 100) = 21.79$, $p < .001$, $\eta_p^2 = .30$, and a significant interaction effect, $F(2, 100) = 3.88$, $p = .024$, $\eta_p^2 = .07$. When the area with the highest noise was in the center of the scatterplot (rather than at the left or right), the subjective fit estimate declined. For the main effects, all pairwise comparisons (Bonferroni-corrected) were significant, $p < .001$. The interaction effect was rather weak but indicated that the differences between areas of highest noise were stronger for high overall noise. The difference between highest noise in Area 1 and highest noise in Area 2 was 5.8 for high overall noise and 3.79 for low overall noise. The difference between highest noise in Area 2 and highest noise in Area 3 was 7.86 for high overall noise and 4.88 for low overall noise.

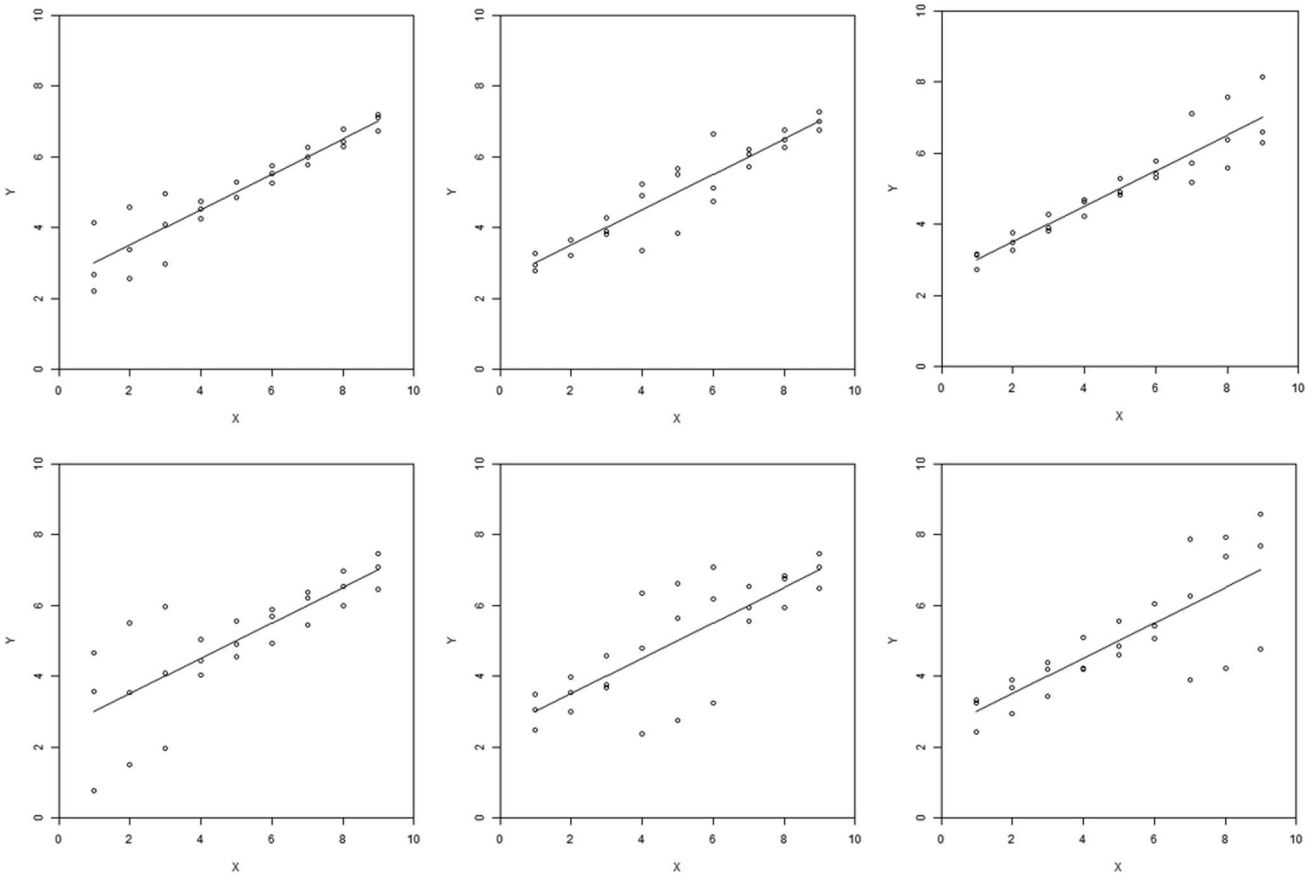

Figure 5. Example scatterplots for Experiment 2. In the left column are scatterplots with the highest noise in Area 1, in the middle column are scatterplots with the highest noise in Area 2, and in the third column are scatterplots with the highest noise in Area 3. Scatterplots in the upper line have a low overall noise (leading to an R^2 of .86), and scatterplots in the lower line have a higher overall noise (leading to an R^2 of .62).

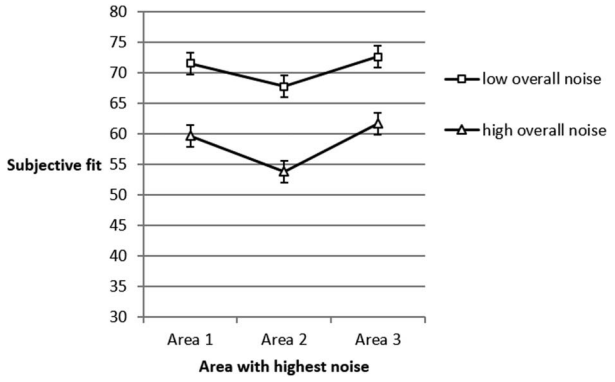

Figure 6. Mean values in rating for each configuration of a scatterplot. Error bars indicate the 95% within subjects CI according to Masson and Loftus (2003) based on the pooled error term.

Discussion

Experiment 2 supported our prediction derived from Experiment 1. We found a consistent pattern across two different levels of overall noise. The fit estimations were lower when the highest noise was located in the center of the scatterplot. The effect size of area of highest noise (η_p^2 = .30) indicates a large effect (cf. Cohen, 1988). In light of the findings in previous work (Tatler, 2007), that the center bias does not disappear when salient image features are placed outside of the center, it is likely that a center bias led to a stronger influence of the point pattern in the center. Despite the center bias documented by Tatler (2007), it seems conceivable that attention was drawn to the periphery by larger noise at least in some trials in our Experiment 2. This would have worked against the effect we nevertheless obtained and reduced our estimate of the extent to which noise in the center is weighed more strongly than in the periphery. Future work involving stronger outlier manipulations and eye tracking might follow up on the question under which conditions outliers in the periphery draw attention and by this increase the impact of noise in the periphery on fit estimates. A second theme to consider for future work concerns the distinction between screen and stimuli. Our scatterplots were placed in the middle of the screen. Bindemann (2010) found evidence for both a screen center bias and a scene/stimuli center bias. Future studies could therefore also vary the position of the graph on the screen.

The results of the experiment have implications for a variety of practical applications. In cases of visual outlier checking, data points in the tails could be less likely or more slowly detected than in the center. Similarly, the quality of checking the assumption of homoscedasticity in regression could strongly vary across different portions of the scatterplot. Furthermore, viewing scatterplots with tails of different variability can be relevant for many cases in the field of visual model assessment and predictive modeling. For example, Wickham et al. (2015) pointed out that model visualization can help to analyze whether the fit in a model is uniformly good or differently good for different regions. In some cases (e.g., Evans et al., 2018), the *x*-axis is related to a temporal dimension. Especially in the case of longer time spans (for example, months or years), it is plausible that the predictions of a model become less accurate over time. For accurate perceptions and conclusions, visualization of such model data patterns requires attention to all regions.

General Discussion

We explored what information people use for estimating fit between model line and data points in scatterplots. Experiment 1 revealed that the perceived fit was rated higher for scatterplots with lower noise and steeper slopes. Thus, participants' visual judgments do not mimic RMSE (which would indicate good fit for a flat line with low noise). Rather, they resemble measures that take variance explained into account. As we learned from Experiment 1 that most of the fixations fell into the center of the scatterplot (center bias), Experiment 2 tested whether noise at central versus peripheral positions in the graph differentially affected the judgment. With this, Experiment 2 addressed a feature that is not related to most statistical fit indices. The results suggested that for the perceived fit, the impact of data points in the center of the scatterplot was higher than the impact of data points in the outer areas.

Taking together, these findings about geometrical properties and gravity of attention and its possible consequences contribute to Brewer (2012) and others' claim to better understand how people use graphs with data and theory to weigh scientific evidence. Additionally, the presented research provides evidence for the existence of and the consequences of the center bias from scene viewing literature (cf. Tatler, 2007), within the perception of scatterplots.

A suggestion for future research is to focus on a more heterogeneous sample, since psychology students might be more familiar with scatterplots and statistical concepts than others. Expertise has been discussed as a confounding factor in visual estimations in scatterplots (cf. Meyer & Shinar, 1992; Strahan & Hansen, 1978).

Across the experiments, we used subjective fit estimation as the dependent variable. The majority of research on subjective impressions in scatterplots has focused explicitly on correlation estimation (e.g., Doherty & Anderson, 2009), and it is possible that some

participants automatically estimated correlation. However, we think it is worthwhile to pursue future research on fit estimations in scatterplots in general. The linear regression line can be used to express a quantitative theory and compare it to data (cf. Kubovy & van den Berg, 2008). The number of possible constellations in arrangement of data points and model (e.g., position and shape) is large. For instance, the linear model does not always have to be the regression line of least squares, and in some applications, one might be interested in the goodness of fit of an already specified model that is tested on new data. Furthermore, some quantitative theories have a curvature shape such as in contexts of learning (e.g., Evans et al., 2018) or forgetting (e.g., Wixted & Ebbesen, 1997). Evaluation of fit might specifically depend on whether a theory systematically overestimates or underestimates the asymptote (cf. Palmeri, 1999). Future research could focus on how individuals estimate fit in those contexts. We understand the presented research as a first step toward that direction.

References

Bergstrom, C. T., & West, J. D. (2018). *Why scatter plots suggest causality, and what we can do about it*. ArXiv. https://arxiv.org/abs/1809.09328

Bindemann, M. (2010). Scene and screen center bias early eye movements in scene viewing. *Vision Research*, *50*(23), 2577–2587. https://doi.org/10.1016/j.visres.2010.08.016

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

Bogen, J., & Woodward, J. (1992). Observations, theories and the evolution of the human spirit, *Philosophy of Science*, *59*(4), 590–611. https://doi.org/10.1086/289697

Brewer, W. F. (2012). The theory ladenness of the mental processes used in the scientific enterprise: Evidence from cognitive psychology and the history of science. In R. W. Proctor & E. J. Capaldi (Eds.), *Psychology of science: Implicit and explicit processes psychology of science: Implicit and explicit processes* (pp. 289–334). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199753628.003.0013

Cleveland, W. S., Diaconis, P., & McGill, R. (1982). Variables on scatterplots look more highly correlated when the scales are increased. *Science*, *216*(4550), 1138–1141. https://doi.org/10.1126/science.216.4550.1138

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge. https://doi.org/10.4324/9780203771587

Doherty, M. E., & Anderson, R. B. (2009). Variation in scatterplot displays. *Behavior Research Methods*, *41*(1), 55–60. https://doi.org/10.3758/BRM.41.1.55

Evans, N. J., Brown, S. D., Mewhort, D. J. K., & Heathcote, A. (2018). Refining the law of practice. *Psychological Review*, *125*(4), 592–605. https://doi.org/10.1037/rev0000105

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/brm.41.4.1149

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

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, *7*(2), 185–207. https://doi.org/10.3758/BF03212979

Kubovy, M., & van den Berg, M. (2008). The whole is equal to the sum of its parts: A probabilistic model of grouping by proximity and similarity in regular patterns. *Psychological Review*, *115*(1), 131–154. https://doi.org/10.1037/0033-295X.115.1.131

Lane, D. M., Anderson, C. A., & Kellam, K. L. (1985). Judging the relatedness of variables: The psychophysics of covariation detection. *Journal of Experimental Psychology: Human Perception and Performance*, *11*(5), 640–649. https://doi.org/10.1037/0096-1523.11.5.640

Lauer, T. W., & Post, G. V. (1989). Density in scatterplots and the estimation of correlation. *Behaviour & Information Technology*, *8*(3), 235–244. https://doi.org/10.1080/01449298908914554

Masson, M. E. J., & Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology*, *57*(3), 203–220. https://doi.org/10.1037/h0087426

Meyer, J., & Shinar, D. (1992). Estimating correlations from scatterplots. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *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

Osborne, J. W., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research and Evaluation*, *8*(2), 1–5. https://doi.org/10.7275/r222-hv23

Palmeri, T. J. (1999). Theories of automaticity and the power law of practice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(2), 543–551. https://doi.org/10.1037/0278-7393.25.2.543

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*(3), 472–491. https://doi.org/10.1037/0033-295x.109.3.472

Reimann, D. (2019). *Visual model fit estimation in scatterplots and distribution of attention: Influence of slope and noise level*. https://doi.org/10.17605/OSF.IO/TG62S

Rensink, R. A. (2017). The nature of correlation perception in scatterplots. *Psychonomic Bulletin & Review*, *24*(3), 776–797. https://doi.org/10.3758/s13423-016-1174-7

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

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358–367. https://doi.org/10.1037/0033-295x.107.2.358

Sarikaya, A., & Gleicher, M. (2018). Scatterplots: Tasks, data, and designs. *IEEE Transactions on Visualization and Computer Graphics*, *24*(1), 402–412. https://doi.org/10.1109/TVCG.2017.2744184

Schnotz, W., & Bannert, M. (2003). Construction and interference in learning from multiple representation. *Learning and Instruction*, *13*(2), 141–156. https://doi.org/10.1016/S0959-4752(02)00017-8

Schunn, C., & Wallach, D. (2005). Evaluating goodness-of-fit in comparison of models to data. In W. Tack (Ed.), *Psychologie der Kognition: Reden und Vorträge anlässlich der Emeritierung von Werner Tack* (pp. 115–154). University of Saarland Press.

Smith, L. D., Best, L. A., Stubbs, D. A., Johnston, J., & Archibald, A. B. (2000). Scientific graphs and the hierarchy of the sciences: A Latourian survey of inscription practices. *Social Studies of Science*, *30*(1), 73–94. https://doi.org/10.1177/030631200030001003

Smith, L. D., Best, L. A., Stubbs, D. A., Archibald, A. B., & Roberson-Nay, R. (2002). Constructing knowledge: The role of graphs and tables in hard and soft psychology. *American Psychologist*, *57*(10), 749–761. https://doi.org/10.1037/0003-066X.57.10.749

Strahan, R. F., & Hansen, C. J. (1978). Underestimating correlation from scatterplots. *Applied Psychological Measurement*, *2*(4), 543–550. https://doi.org/10.1177/014662167800200409

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, *7*(14), 4, 1–20. https://doi.org/10.1167/7.14.4

Wagenmakers, E.-J. (2003). How many parameters does it take to fit an elephant? *Journal of Mathematical Psychology*, *47*(5–6), 580–586. https://doi.org/10.1016/S0022-2496(03)00064-6

Wickham, H., Cook, D., & Hofmann, H. (2015). Visualizing statistical models: Removing the blindfold. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *8*(4), 203–225. https://doi.org/10.1002/sam.11271

Wixted, J. T., & Ebbesen, E. B. (1997). Genuine power curves in forgetting: A quantitative analysis of individual subject forgetting functions. *Memory & Cognition*, *25*(5), 731–739. https://doi.org/10.3758/BF03211316

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

History
Received December 4, 2019
Revision received August 19, 2020
Accepted October 12, 2020
Published online December 4, 2020

Open Data
The raw data underlying the findings reported in the article as well as the materials are available at the public data repository https://osf.io/tg62s/

ORCID
Daniel Reimann
https://orcid.org/0000-0003-4687-8858

Daniel Reimann
Department of Psychology
FernUniversität in Hagen
Universitätsstraße 33
58097 Hagen
Germany
daniel.reimann@fernuni-hagen.de

Appendix

Further Information on Results

Table A1. Means (*M*), Standard Deviations (*SD*s), and reliability for the ratings for Experiment 1

Slope	Noise *SD* = 1 *M* (*SD;* α)	Noise *SD* = 2 *M* (*SD;* α)	Noise *SD* = 3 *M* (*SD;* α)
22.5°	65.96 (14.74; .86)	40.19 (19.39; .91)	27.35 (21.15; .95)
45°	75.15 (9.48; .72)	49.50 (19.63; .91)	33.02 (19.90; .90)
67.5°	88.11 (5.89; .79)	70.00 (11.71; .82)	57.51 (15.70; .87)

Note. $N = 25$. α = Cronbach's alpha. This table refers to the values in Figure 2. It shows the mean value in estimation of fit between model and data for each of the nine experimental conditions. Since we created four parallel versions for each condition, we could provide information about the reliability.

Table A2. Means (*M*), Standard Deviations (*SD*s), and reliability for the ratings for Experiment 2

Overall noise	Highest noise Area 1 *M* (*SD;* α)	Highest noise Area 2 *M* (*SD;* α)	Highest noise Area 3 *M* (*SD;* α)
Low	71.58 (15.37; .93)	67.80 (15.12; .95)	72.68 (14.54; .92)
High	59.15 (19.28; .96)	53.89 (18.25; .95)	61.75 (16.46; .95)

Note. $N = 51$. α = Cronbach's alpha. This table refers to the values in Figure 6. It shows the mean value in estimation of fit between model and data for each of the six experimental conditions. Since we created six parallel versions for each condition, we could provide information about the reliability.

