A Novel Technique to Facilitate More Accurate Correlation Perception in Scatterplots

Gabriel Strain[*] Andrew J. Stewart[†] Paul Warren[‡] Caroline Jay[§]

The University of Manchester

Abstract

Viewers consistently underestimate correlation in positively correlated scatterplots. We use a novel point size manipulation to correct for this bias. In a high-powered and fully reproducible study, we demonstrate that decreasing the size of a point in a scatterplot as a function of its distance from the regression line is able to correct for a systematic perceptual bias long present in the literature. We suggest the implementation of our technique when designing scatterplots that aim to communicate correlation.

Index Terms: Human-centered computing—Visualization—Empirical studies in visualization—; Human-centered computing—Human computer interaction (HCI)—Empirical studies in HCI—

1 Introduction

Scatterplots, utilized in scientific communication for a variety of tasks, are some of the most widely used and studied data visualizations. Viewers interpret them in similar ways [11], and they are simple enough to be easily studied while providing important insights into visualization design, human-computer interaction, and human perception. In a previous study [25], we showed that a novel point contrast manipulation, in which the contrast of a certain scatterplot point was reduced as the size of that point's residual increased, could be used to partially correct for a systematic correlation underestimation bias present in the literature [4–6, 13, 14, 17, 24]. We suggested that this was due to a narrowing of the width of the perceived probability distribution of a plot brought about by the lower contrast (and therefore lower point-salience and higher spatial uncertainty) in those outer areas. In that study we tested linear, non-linear, and non-linear inverted functions relating point contrast to residual size, finding that the non-linear function produced the most accurate estimates of correlation, and that the non-linear inverted produced the least accurate. In the present study we use the same equations to manipulate point size.

1.1 Scatterplots and Correlation

Scatterplots have been widely studied, especially as mediums for the communication of correlation (see [25] for a review of the history of this work). Previous literature has found evidence for a pronounced underestimation in judgements of correlation in positively correlated scatterplots, especially between $0.2 < r < 0.6$. The nature of this investigation has varied, ranging from direct estimation, to discriminative judgement, bisection, and staircase tasks. As in our previous work, we use the direct estimation paradigm owing to its simplicity and its suitability to online experimentation. This renders the judgements we collect comparative by nature, although such work does allow us to inform design guidelines as well as human perception. It is our duty as visualization designers to ensure that the messages we are trying to communicate are being interpreted as accurately as possible by viewers. To achieve this, we must understand human perception, apply that understanding to design, and test those designs in rigorous empirical studies.

1.2 Point Size

Point contrast is not the only available visual feature that might be used to influence viewers' perceptions of the width of a probability distribution, neither is it the only visual feature of a scatterplot that we can exploit. While contrast adjustments have been used extensively to solve issues of overplotting and clutter in scatterplots [3, 16], there is no established use for varying point size. Common sense dictates that scatterplots visualizing larger datasets inherently require their points to be smaller to prevent obfuscation of the data, but to our knowledge there is little testing of the impact of point size on correlation perception. Studies have found invariance in the bias and variability of correlation perception with regards to changing point sizes, but these have been low-powered [22, 23]. From the wider literature there is evidence that larger points are more salient [9], can bias judgements of point position more strongly than point contrast can [10], and can result in faster reaction times to peripherally presented stimuli [8]. In addition, smaller stimuli are associated with greater levels of spatial uncertainty [1], and if this is driving the reduction in bias we saw in our previous work [25], we would expect a similar effect when point size is used instead of point contrast.

1.3 Hypotheses

We hypothesize that correlation estimates will be most accurate when viewers are presented with the non-linear size decay condition, and will be least accurate when presented with the non-linear inverted size decay condition. We thereby present a single experiment study in which we demonstrate that the use of a non-linear size decay function relating to the residuals of points on scatterplots can be employed to correct for a systematic underestimation of correlation by viewers. We find no evidence for effects of graph literacy or training. The effect we observe here is much stronger, both with regards to effect size and in terms of the observed reduction in error, than that observed in our previous study [25]. We suggest that this function can be used to facilitate more accurate correlation perception in scatterplots, and provide exciting future avenues for the continuation and refinement of these techniques. Ethical approval was granted by the University of Manchester's Computer Science Departmental Panel (Ref: 2022-14660-24397).

2 Methodology

The experiment was conducted according to the principles of open and reproducible research. All data and analysis code are available at `https://github.com/gjpstrain/size_and_scatterplots`. This repository contains instructions for building a docker image to fully reproduce the computational environment used, allowing for full replications of stimulus generation, analyses, and the paper itself. The experiment was pre-registered with the OSF (`https://osf.io/k4gd8`).

[*] Gabriel.Strain@manchester.ac.uk

[†] Andrew.J.Stewart@manchester.ac.uk

[‡] Paul.Warren@manchester.ac.uk

[§] Caroline.Jay@manchester.ac.uk

2.1 Participants

150 participants were recruited using the Prolific.co platform. Normal to corrected-to-normal vision and English fluency were required for participation. As in our previous work [25], and in accordance with previously published guidelines [19], participants were required to have completed at least 100 studies on Prolific, and were required to have a Prolific score of at least 100, indicating acceptance on at least 100/101 previously completed studies. Participants who took part in any of our previous studies were prevented from participating, and participants were only permitted to complete the experiment on a desktop or laptop computer.

Data were collected from 164 participants. 14 failed more than 2 out of 6 attention checks, and, as per pre-registration stipulations, were rejected from the study. Data from 150 participants was included in the analysis (48% male, 50% female, and 2% non-binary). Mean age of participants was 29.6 (*SD* = 8.5). Mean graph literacy score was 21.77 (*SD* = 4.29) out of 30. The average time taken to complete the experiment was 39 minutes (SD = 14 minutes).

2.2 Stimuli

The data used to generated the scatterplots in the current study were identical to that used previously [25]. Scatterplots were generated based on 45 uniformly distributed *r* values between 0.2 and 0.99. Scatterplot points were generated based on bivariate normal distributions with standard deviations of 1 in each direction. Each scatterplot had a 1:1 aspect ratio, was generated as a 1200 x 1200 pixel .png image, and was scaled up or down according to the participant's monitor. See Sect. 2.3 for a more detailed discussion of precise point sizes and dot pitch in crowd-sourced experiments.

As in our previous study [25], we used equation 1 to map residuals to point sizes in three of our conditions. We additionally used a scaling factor of 4 and a constant of 0.2 to achieve a minimum on-screen point size of 12 pixels, which is consistent with the point size on a 1920 x 1080 monitor for both experiments in [25].In our fourth condition, which we refer to as *standard size*, point size was uniformly set to be consistent with the point size in our previous studies. Scripts detailing scatterplot and mask generation can be found in the item preparation folder in the repository linked below.

$$point_{size} = 1 - b^{Residual} \qquad (1)$$

2.3 Dot Pitch and Crowdsourced Experiments

In our previous study [25], we had no way of obtaining dot pitch or participant to monitor distance due to the online, crowdsourced nature of the experiments. Since then we have adopted a method for obtaining the height of a participant's monitor in inches [18]; participants are asked to hold up a standard size credit/debit/ID card up to the monitor, and then to resize a corresponding image until it matches the physical size of the card. These cards have a universal standard size (ISO/IEC 7810 ID-1), which when combined with the monitor resolution obtained from PsychoPy [20] and assuming a widescreen 16:9 aspect ratio, allows us to infer dot pitch and therefore the physical size of the points in our experiment. Mean dot pitch was 0.33mm, ($SD = 0.06$), corresponding to a physical size on the screen of 4.32mm for the smallest points displayed. See Sect. 3 for analyses including dot pitch as a predictor.

2.4 Visual Threshold Testing

It is key that our manipulation does not remove data from the scatterplot, thus, in order to test that all our points were visible across a range of viewing contexts and on a range of apparatus, we included visual threshold testing prior to the experimental items in the study. Participants were shown six scatterplots and were asked to enter in a text box how many points were being displayed. The points were the same size as the smallest points used in the experimental materials. 5% of participants were correct on 5 out of 6 visual threshold questions, while 95% were correct on 6 out of 6. It should be noted that those participants scoring 5/6 did not answer incorrectly, rather they did not answer at all for a particular question, which is suggestive of a mis-click or an initial misunderstanding of the task. Regardless, we consider these results to be indicative of a sufficient level of point visibility.

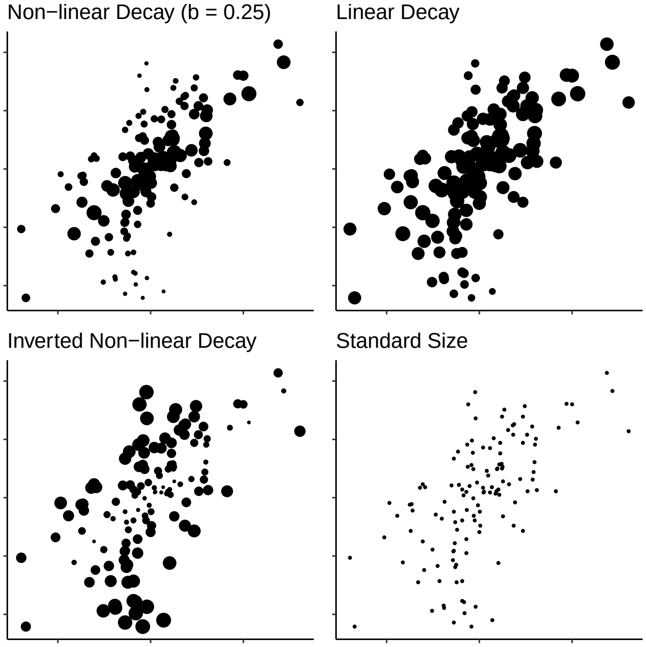

Figure 1: Four levels of the point size decay condition, demonstrated with an *r* value of 0.6

2.5 Design

The experiment used a fully repeated measures, within-participants design, with each participant seeing and responding to each of the 180 scatterplots in a randomized order. There were four scatterplots for each of the 45 *r* values corresponding to the four levels of the size decay condition, examples of which can be see in Figure 1. Everything needed to run the experiment, including code, materials, instructions, and scripts, is hosted at `https://gitlab.pavlovia.org/Strain/exp_size_only`.

2.6 Procedure

Each participant was shown the participant information sheet (PIS) and provided consent through key presses in response to consent statements. They were asked to provide their age in a free text box, and their gender identity. Participants then completed the 5-item Subjective Graph Literacy test [7], followed by the visual threshold testing described above. Participants then completed the screen scaling task described in Sect. 2.3. Participants were given instructions, and then shown examples of *r* = 0.2, 0.5, 0.8, and 0.95. Sect. 4.1 includes a discussion of the potential effects of viewing these examples. Two practice trials were given before the experiment began. Participants worked through a series of 180 trials and were asked to use a slider to estimate the correlation shown in the scatterplot. Visual masks preceded each plot. Interspersed were six attention check trials which asked participants to set the slider to 1 or 0 and ignore the scatterplot.

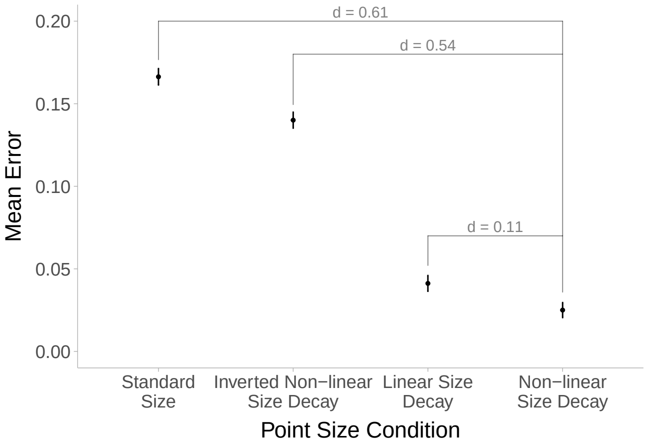

Figure 2: Mean error in correlation estimates across the four conditions, with 95% confidence intervals. Effect sizes between standard size and other conditions in Cohen's d are also displayed

Table 1: Contrasts between the four levels of the size decay condition.

Contrast	Z.ratio	p.value
Non-linear Decay: Linear Decay	-3.99	<0.001
Non-linear Decay : Inverted Non-linear Decay	-20.57	<0.001
Non-linear Decay : Standard Size	-15.41	<0.001
Linear Decay : Inverted Non-linear Decay	-16.86	<0.001
Linear Decay : Standard Size	-11.96	<0.001
Inverted Non-linear Decay : Standard Size	-3.63	0.002

3 Results

All analyses were conducted using R (version 4.3.0 [21]). Models were built using the **buildmer** (version 2.8 [26]) and **lme4** (version 1.1-32 [2]) packages, with size decay condition being set as the predictor for participants' errors in correlation estimates.

Mean errors in correlation estimates for the four size decay conditions can be seen in Figure 2. A likelihood ratio test revealed that the model including size decay condition as a predictor explained significantly more variance than a null model ($\chi^2(3)$ = 205.35, p < .001). This model has random intercepts for items and participants. The effect here is driven by participants' errors being lower for scatterplots with the non-linear size decay manipulation than for all other conditions, for error being lower for scatterplots with linear size decay than for plots with inverted non-linear decay or standard size, and for errors being higher for scatterplots with standard size than for plots with inverted non-linear decay.

Testing for contrasts between the four levels of the size decay condition was performed with the **emmeans** package (version 1.8.5 [15]), and are shown in Table 1. The **EMAtools** (version 0.1.4 [12]) package was used to calculate effects sizes in Cohen's d, the results of which can be seen in Figure 2. The largest effect size we found was 0.61 when comparing the non-linear size decay and standard size decay conditions. This is significantly higher than any of the effects sizes we found in our previous work.

In addition, we find no significant difference between the experimental model and another including graph literacy as a fixed effect ($\chi^2(1)$ = 0.45, p = .502), suggesting the effect we found was not driven by differences in graph literacy.

Figure 4 shows how participants' mean errors in correlation estimates change with the objective r value, plotted separately for each size decay condition. Note the close-to-zero average errors present in the non-linear size decay condition.

We employed a method for obtaining a measurement of dot pitch from each participant. While Sect. 2.4 provides evidence that participants had no problems perceiving all the points shown on the scatterplots, there may be some other facet of using a larger or smaller monitor with a higher or lower resolution that could have affected the estimates participants gave. To check this, we built a model including the dot pitch measurement as a fixed effect. Comparing this to the experimental model revealed a significant effect of dot pitch ($\chi^2(1)$ = 4.44, p = .035). There was no interaction between size decay condition and dot pitch, with a decrease in dot pitch of 0.1 resulting in a decrease in estimated correlation of .03. While significant, we do not consider this effect large enough to warrant further discussion.

4 Discussion

As can be seen in Figure 4, participants' errors in correlation estimates were significantly lower when they were presented with the non-linear size decay condition (see Figure 1) compared to when they were presented with all other conditions, providing support for our first hypothesis. We found no support for our second hypothesis, that participants' estimates would be least accurate in the inverted non-linear size decay condition. Errors in this condition were indeed significantly higher than for the other two size decay conditions, but were significantly lower than the error with the standard size condition.

The mean error in correlation estimation for the non-linear size decay condition used in the present study was 0.025, while the equivalent condition in the second experiment of our previous study, which used the same equations applied to contrast, resulted in a mean error of 0.086 [25]. Taken together, this is evidence that point size is a much stronger channel for the manipulation of perceived correlation in scatterplots than point contrast. If the effects we have found here and in our previous work are being driven by increased uncertainty in the outer regions of the plots, that we have found a large effect of point size manipulations is congruent with research showing clear influences of stimulus size on perception and uncertainty [1, 8, 10]. Contrary to this, the literature linking contrast to perceptual uncertainty is sparse. Unlike our previous work, in which the standard deviations of errors for most conditions became smaller as the objective r value increased, participants' distributions of standard deviations of correlation estimates remained mostly constant. This is unexpected, as previous work, including our own, finds precision in r estimation to increase as the objective r value increases. Given that we found this in our work manipulating point contrast [25], and its robustness in the literature, this result is surprising. We suggest that this is due to the nature of the stimuli. At high values of r there is a large amount of overlap between points with the non-linear, non-linear inverted, and linear size decay conditions (see examples in the item preparation folder in the repository linked above). It may be that this is itself producing greater uncertainty and causing an absence of the increased precision we would expect to see at higher r values. While the visual character of the scatterplots in the aforementioned conditions can account for the absence of higher precision at higher r values, the same cannot be said for the standard size condition. Aside from the inverted non-linear decay condition in experiment two of our prior work [25], the finding that precision increased with r was robust. Its absence here is curious given that the standard size decay condition in the present study is identical to the full contrast conditions in our previous work. Taken together, this suggests that there is something particular about the scatterplots in the present study that is causing this. Relying on relative judgements means the interplay between scatterplots with different visual features must be accounted for within a particular experiment. Here, this interplay has resulted in the absence of this effect, albeit further testing would be required for a more concrete explanation.

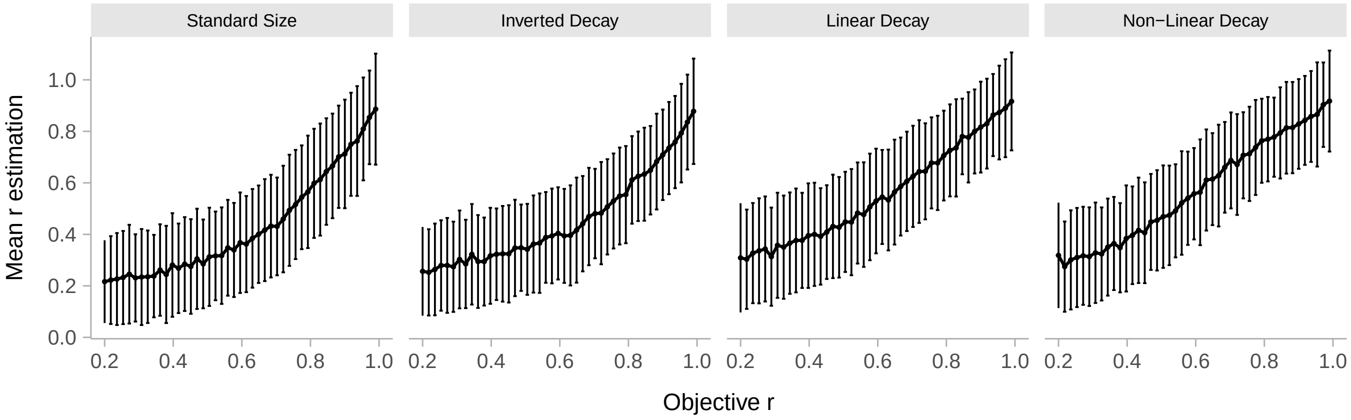

Figure 3: Participants' mean *r* estimates plotted against the objective *r* value separately for each size decay condition.

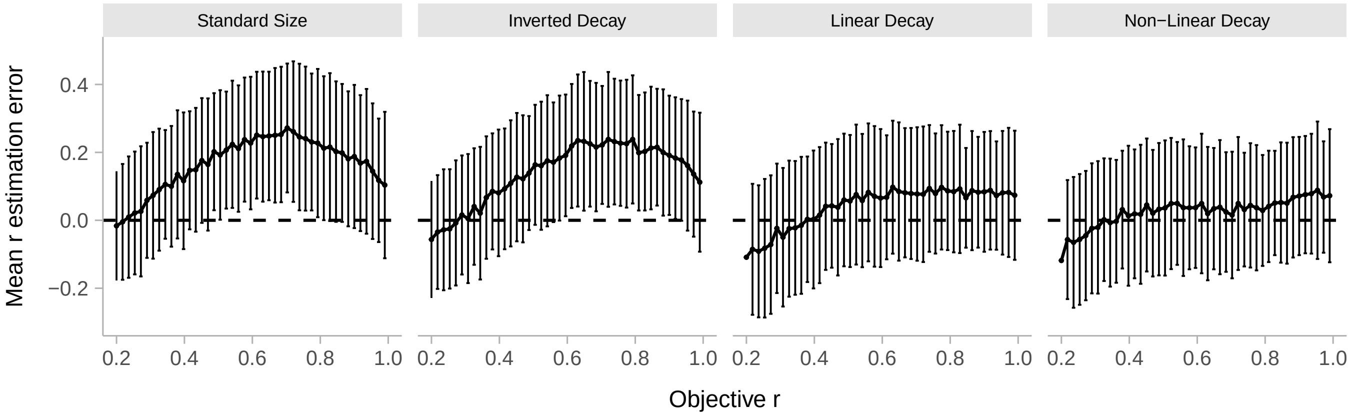

Figure 4: Participants' mean errors in *r* estimates plotted against the objective *r* value separately for each size decay condition. The dotted horizontal line represents perfectly accurate estimation.

The lack of support for our second hypothesis was surprising, although it should be noted that the difference between the inverted non-linear condition and the standard size condition was small (see Figure 2). The shape of the errors in correlation estimates (see Figure 4) is also very similar to that of the standard size decay condition. It would seem then that despite the size channel being more powerful than contrast with regards to correcting for the underestimation of correlation, it is weaker than the contrast channel at producing the opposing effect. In our previous work we suggested that contrast manipulations could be used to correct for the *overestimation* of the correlation of negatively correlated scatterplots; we would not suggest the use of the size channel for this given the results here.

To conclude, we recommend the use of the non-linear size decay condition described here when designing scatterplots optimized for correlation perception.

4.1 Training

Before the experiment, participants viewed plots for a minimum of eight seconds with examples of r = 0.2, 0.5, 0.8, and 0.95. This was to account for any potential unfamiliarity with scatterplots present in the samples that we recruited; this risk is inherent in recruiting from lay populations, but we would argue is acceptable given it leads to more generalisable and broadly applicable findings. To test whether this training had an effect on correlation estimation, we built a model including session half as a predictor. Comparing this to the original model revealed no significant effect ($\chi^2(1) = 1.28$, $p = 0.26$), suggesting that having more recently viewed the example plots did not have an effect on participants' estimates of correlation.

4.2 Limitations

The data we have gathered is inherently comparative. Despite confirming a method of obtaining dot pitch, we still have no method of obtaining head-to-monitor distances. Taken together, these aspects of our experiment prevent us from making concrete psychophysical conclusions, but instead allow for findings that are rigorous to different viewing contexts that we argue are of particular importance for the HCI and design audiences. It may be that a high level perceptual phenomenon is responsible for the effects we have seen here; investigating this is beyond the scope of the current study and does not negate our findings.

4.3 Future Work

At present, we have confirmed the potential for both point contrast and size manipulations to influence participants' perceptions of correlation in scatterplots, each to varying degrees. It is also clear that these manipulations are not necessary, and may even be making perception worse, at very low and high values of *r*. Our future work will therefore take two directions; we will investigate the effect of manipulating both point size and contrast on correlation estimation, and we will introduce a parameter to control the strength of this family of manipulations according to the objective *r* value itself.

Acknowledgments

This work was supported by funding from the University of Manchester Department of Computer Science and Division of Psychology, Communication and Human Neuroscience.

References

[1] D. Alais and D. Burr. The Ventriloquist Effect Results from Near-Optimal Bimodal Integration. *Current Biology*, 14(3):257–262, Feb. 2004. doi: 10.1016/j.cub.2004.01.029

[2] D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01

[3] E. Bertini and G. Santucci. Quality Metrics for 2D Scatterplot Graphics: Automatically Reducing Visual Clutter. In A. Butz, A. Krüger, and P. Olivier, eds., *Smart Graphics*, Lecture Notes in Computer Science, pp. 77–89. Springer, Berlin, Heidelberg, 2004. doi: 10.1007/978-3-540-24678-7_8

[4] P. Bobko and R. Karren. The Perception of Pearson Product Moment Correlations from Bivariate Scatterplots. *Personnel Psychology*, 32(2):313–325, 1979. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1744-6570.1979.tb02137.x. doi: 10.1111/j.1744-6570.1979.tb02137.x

[5] W. S. Cleveland and R. McGill. Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984. Publisher: [American Statistical Association, Taylor & Francis, Ltd.]. doi: 10.2307/2288400

[6] C. E. Collyer, K. A. Stanley, and C. Bowater. Psychology of the Scientist: LXIII. Perceiving Scattergrams: Is Visual Line Fitting Related to Estimation of the Correlation Coefficient? *Perceptual and Motor Skills*, 71(2):371–378E, Oct. 1990. Publisher: SAGE Publications Inc. doi: 10.2466/pms.1990.71.2.371

[7] R. Garcia-Retamero, E. T. Cokely, S. Ghazal, and A. Joeris. Measuring Graph Literacy without a Test: A Brief Subjective Assessment. *Medical Decision Making*, 36(7):854–867, 2016. Publisher: SAGE Publications Inc STM. doi: 10.1177/0272989X16655334

[8] G. R. Grice, L. Canham, and J. M. Boroughs. Forest before trees? It depends where you look. *Perception & Psychophysics*, 33(2):121–128, Mar. 1983. doi: 10.3758/BF03202829

[9] C. Healey and J. Enns. Attention and Visual Memory in Visualization and Computer Graphics. *IEEE Transactions on Visualization and Computer Graphics*, 18(7):1170–1188, July 2012. Conference Name: IEEE Transactions on Visualization and Computer Graphics. doi: 10.1109/TVCG.2011.127

[10] M.-H. Hong, J. K. Witt, and D. A. Szafir. The Weighted Average Illusion: Biases in Perceived Mean Position in Scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2021. doi: 10.1109/TVCG.2021.3114783

[11] M. Kay and J. Heer. Beyond Weber's Law: A Second Look at Ranking Visualizations of Correlation. *IEEE transactions on visualization and computer graphics*, 22, Sept. 2015. doi: 10.1109/TVCG.2015.2467671

[12] E. Kleiman. *EMAtools: Data Management Tools for Real-Time Monitoring/Ecological Momentary Assessment Data*, 2021. R package version 0.1.4.

[13] D. Lane, C. Anderson, and K. Kellam. Judging the Relatedness of Variables. The Psychophysics of Covariation Detection. *Journal of Experimental Psychology: Human Perception and Performance*, 11:640–649, Oct. 1985. doi: 10.1037/0096-1523.11.5.640

[14] T. W. Lauer and G. V. Post. Density in scatterplots and the estimation of correlation. *Behaviour & Information Technology*, 8(3):235–244, June 1989. Publisher: Taylor & Francis. doi: 10.1080/01449298908914554

[15] R. V. Lenth. *emmeans: Estimated Marginal Means, aka Least-Squares Means*, 2023. R package version 1.8.5.

[16] J. Matejka, F. Anderson, and G. Fitzmaurice. Dynamic Opacity Optimization for Scatter Plots. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 2707–2710. ACM, Seoul Republic of Korea, Apr. 2015. doi: 10.1145/2702123.2702585

[17] J. Meyer and D. Shinar. Estimating Correlations from Scatterplots. *Human Factors*, 34(3):335–349, June 1992. Publisher: SAGE Publications Inc. doi: 10.1177/001872089203400307

[18] W. L. Morys-Carter. ScreenScale, May 2021. https://doi.org/10.17605/OSF.IO/8FHQK.

[19] E. Peer, D. Rothschild, A. Gordon, Z. Evernden, and E. Damer. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, Sept. 2021. doi: 10.3758/s13428-021-01694-3

[20] J. Peirce, J. R. Gray, S. Simpson, M. MacAskill, R. Höchenberger, H. Sogo, E. Kastman, and J. K. Lindeløv. PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 51(1):195–203, Feb. 2019. doi: 10.3758/s13428-018-01193-y

[21] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023.

[22] R. Rensink. Invariance of Correlation Perception. vol. 12, pp. 433–433, May 2012. doi: 10.1167/12.9.433

[23] R. A. Rensink. On the Prospects for a Science of Visualization. In W. Huang, ed., *Handbook of Human Centric Visualization*, pp. 147–175. Springer New York, New York, NY, 2014. doi: 10.1007/978-1-4614-7485-2_6

[24] R. F. Strahan and C. J. Hansen. Underestimating Correlation from Scatterplots. *Applied Psychological Measurement*, 2(4):543–550, Oct. 1978. Publisher: SAGE Publications Inc. doi: 10.1177/014662167800200409

[25] G. Strain, A. J. Stewart, P. Warren, and C. Jay. The Effects of Contrast on Correlation Perception in Scatterplots. *International Journal of Human-Computer Studies*, 176:103040, Aug. 2023. doi: 10.1016/j.ijhcs.2023.103040

[26] C. C. Voeten. *buildmer: Stepwise Elimination and Term Reordering for Mixed-Effects Regression*, 2023. R package version 2.8.