**Abstract**

**Introduction**

Data visualisations are presentations or representations of data that relies on peoples visual perception to amplify understanding of data. Further, they are a ubiquitous methodology that provides people with a mechanism to understand complex data in visual-image form (Hehman & Xie, 2021). Moreover, data visualisations are present in the public, private, academic, and social media sectors (Wilke, 2019). A core research goal is the enhancement of data visualisations to allow for individuals to better understand data. Within this paper, we will build on prior literature and methodology to ascertain how certain modifications to data visualisations can enhance understanding. Specifically, we will focus on scatterplots and modify the point encoding feature of size using Strain and colleagues (2023) novel transformation to enhance estimation of correlation scatterplots.

**Making good data visualisations**

Tufte's (1982, 2001) seminal work on data visualisations offered some key axioms of how to produce data visualisations that are effective insofar as they embody what a data visualisation should be. They should show the data, show as much of the data points as possible, and reveal the “truth” of the data by avoiding distortions, fabrications, or errors (Tufte, 2001). Further, Hehman and Xie (2021) build upon these axioms and posit three related guiding philosophies. First, like Tufte, they argue for information richness: showing as much of the data as possible and telling the truth. Hehman and Xie (2021) argue that using summary visualisations, e.g., box plots, that only examine a central tendency can mislead viewers. In support, Weissgerber et al. (2019) suggest that box plots conceal data as they aggregate and do not show the underlying data distribution. To improve this, recent R packages have allowed researchers to combine plots. A common and effective plot combination for boxplots is to combine then with violin plots and jitter plots (Casals & Daunis-i-Estadella, 2023; Cui, 2020); this produces a plot that shows the aggregate, underlying data, and distribution (Kassambara, 2019; Stewart, 2022). Further, raincloud plots similarly produce this effective combination of plot factors (Stewart, 2019).

Hehman and Xie's (2021) second philosophy is minimalism. They argue that having excessive amnesties, e.g., three-dimensional graphs where two-dimensional graphs suffice, lower reader comprehension (Wilke, 2019). Moreover, unnecessary additions like shadowing or mirroring effects should be avoided. Finally, Hehman and Xie (2021) argue that using colour effectively is essential. Their first consideration is that approximately 5% of humans and closer to 10% of males are colour blind (Neitz & Neitz, 2011). The most common colour blindness is red-green also known as deuteranopia; this causes issues seeing and distinguishing between red, green, and yellow (Neitz & Neitz, 2011). So, to be inclusive, understanding what colours to avoid is essential. Further, in many journals and scientific communication mediums, grey scale (graph categories being distinguished by shades of grey) is common (Holte & Ferraro, 2023). Thus, the visualisations researchers make should be meaningful in grey scale and not rely on needing specific colours, e.g., green.

**Studying scatterplots**

Scatterplots, one of the most widely used data visualisation techniques, visually represent bivariate data across x and y cartesian coordinates (Doherty et al., 2007). Scatterplots are worthy of studying for, among other things, three key reasons. First, they are widely used and thus represent practical significance; if researchers can improve scatterplots whereby people estimate or understand them better, then a wide array of data can be better understood (Gleicher et al., 2013). Second, scatterplots are at the ideal threshold of complexity. They are simpler than other visualisations like multidimensional data visualisations, e.g., violin-box plots, but are complex enough to show useful insights, e.g., correlations between two variables of interest (Rensink & Baldridge, 2010). Thus, they are easily studied and provide beneficial insights into data visualisation perception. Finally, scatterplots are excellent candidates for data visualisation research because they are easily modified, for example, the size, colour, position, and opacity of scatterplot points can be easily modified to enhance or reduce perceptual estimation (Sarikaya & Gleicher, 2018). Overall, they are a good choice of data visualisation for rigorous experimental psychology research.

**A primer on correlation perception research**

Much research has been conducted on scatterplots and researchers have provided useful insights into how they are perceived. While scatterplots can be used for several goals, e.g., unsupervised machine learning such as cluster analysis and assumption tests like outlier detection and qq-plots, the most common reason for using scatterplots is to assess whether a correlation between two variables is present (Sarikaya & Gleicher, 2018). Prior research has questioned what data visualisation is most apt at representing correlations. Two separate studies using different mathematical models agreed that scatterplots are the best data visualisation for displaying correlations (Harrison et al., 2014; Kay & Heer, 2016). While several statistical methods exist, e.g., Spearman’s correlation and Kendall’s correlation, the most used correlation method is the Pearson product-moment correlation coefficient.

***Pearson product-moment correlation coefficient***

The Pearson product-moment correlation coefficient which we will denote as Pearson’s r, is a statistical method that assesses the linear correlation between two sets of data. For example, popular correlation research shows strong associations between googling cancer diagnoses and dying of cancer (Wehner et al., 2017), while one study suggests no correlation between attractiveness and intelligence (Mitchem et al., 2015). An axiomatic but necessary point that must be stated is that correlation or association between two variables does not and cannot tell us about causation.

The formula to work out Pearson’s r is below whereby and denotes the values of the x and y variables in the studied sample and and denote the mean of these variables. Pearson’s r provides an r value that ranges between 0 and the absolute value of 1 denoted as |1|. |1| can range from -1 to 1. Thus, the r value can assess positive and negative correlations.

**Seminal Correlation perception research (mid-to-late twentieth century)**

***Schools of thought***

To understand contemporary correlation perception research, it can be beneficial to examine seminal or pioneering work. Feyerabend (2020) argues that it is important to understand the historical works within a field to fully comprehend the subject. First, seminal work in correlation perception was distinguished between two schools of thought: axiomatisers and scalers. Axiomatisers create scales based on measurement and mathematical theories. Fundamentally, axiomatisers construct qualitative conditions to justify scales (Krantz et al., 1971; Krantz, 1972; Krantz & Tversky, 1971; Luce et al., 1990; Suppes et al., 1989). On the other hand, scalers are interested in participants assigning numbers to scales, e.g., rating correlation strength on a scale (Ellermeier & Faulhammer, 2000; Stevens, 1951, 1975, 1957). As the present study is focused on getting participants to assign an r value to a scatterplot with a scale, the focus of the next sections will primarily be on research within the scaler paradigm.

***Measuring correlation perception***

Within early scaling work, two types of measurements were predominantly used: discrimination and estimation. Discrimination is typically when participants are asked to compare two or more stimuli, e.g., two side-by-side graphs (Pollack, 1960). Estimation is where participants are asked to estimate a particular stimulus, e.g., what the r value is (Strahan & Hansen, 1978). As the present study adopts an estimation paradigm, we will focus on estimation research. Early work asked participants to make discriminative and comparative judgements between two or more graphs; they found that as the maginitude of the r value increased, judgements were more accurate (Pollack, 1960).

***Core findings***

Additionally, seminal findings suggested that participants can rapidly extract relatively accurate correlation information, e.g., large or small correlation (Bobko & Karren, 1979; Pollack, 1960). Importantly, research found that expertise did not influence r value estimation performance (Meyer et al., 1997; Meyer & Shinar, 1992). Further, an important early finding was that participants consistently underrate positive r values 0 < r < 1 in estimation studies (Bobko & Karren, 1979; Lane et al., 1985). Moreover, several studies found that this systematic effect was pronounced when the r value is .2 < r < .6 (Bobko & Karren, 1979; Cleveland et al., 1982; Cleveland & McGill, 1984). Overall, seminal research found a systematic effect that demonstrated the need to improve correlation visualisations so that participants underestimate r values less (Elliott, 2021).

**Contemporary Correlation perception research (twenty-first century*)***

While seminal findings of correlation perception were useful, a modern criticism of this research is that they are antiquated and had design flaws not shared by contemporary psychophysics research (Elliott, 2021). The findings of Doherty et al. (2007) contrasted with prior findings of a systematic underestimation of correlation values. They suggested that people overestimate midrange correlations and underestimated large correlations. Regardless, this finding still demonstrated the need to improve visualisation design due to the systematic over-and-underestimation of correlation values.

A resurgence in adequate methodology and studies was started by Rensink & Baldridge (2010) where they found just-noticeable-differences (JND) of participant’s discrimination judgements. Further, Rensink (2012) found no direct estimation differences between scatterplots when manipulating colour, size, shape, and brightness among others. Importantly, contemporary research has developed and applied mathematical laws that compute the mechanisms of correlation perception (Rensink, 2016, 2017).

**Mechanisms of correlation perception**

***Laws of correlation perception***

***Weber’s law***

The first formula that can be used to demonstrate correlation perception is Weber’s law. Research has suggested that the relationship between perception of differences and objective differences can be understood linearly (Harrison et al., 2014; Rensink, 2012). Within this formula is the differential perceptual change, is the change in stimulus, and is the overall correlation or the stimulus. , also known as a Weber fraction, can be derived experimentally (Harrison et al., 2014). Altogether, these parameters form a Weber model that models the perception of correlations in scatterplots.

***Rensink’s instance of Weber’s law***

Rensink (2017) further developed Weber’s law to understand discrimination to compute the value of two scatterplots. Here, K describes an instance of the Weber’s fraction, is an instance of bias or the offset in perceptual discrimination, and is + 0.5 \* .

***Rensink’s instance of Fechner’s law***

Importantly, Rensink (2016, 2017) were able to demonstrate that Fechner’s law, related to Weber’s law, could be used to understand perceptual estimation. Here, best describes the offset in perceptual estimation between the perceived estimation and objective value. Importantly Rensink et al found that the Fechner assumption of = to be true, systematically connecting estimation and discrimination. Further, this assumption suggests that scatterplot studies of different distributions and number of dots facilitate similar performances (Ip et al., 2021). Therefore, base level changes, e.g., comparing large dots to small dots or blue to red dots do not change performance (Rensink, 2012). Overall, this suggests that more novel modifications should be studied.

***Further Correlation perception drivers***

Importantly, while the laws described can model correlation estimation, other factors that drive perceptual ability have been proposed. First, Meyer and Shinar (1992) demonstrated that subjective estimations of correlation strength are a function of the deviation from a perfect correlation of 1. This suggests that individuals intuitively use the mean distance from the regression line to estimate correlation values (Meyer et al., 1997; Meyer & Shinar, 1992).

Further, as stated previously, most studies suggest that people underestimate correlation strength within the .2 to .6 range (e.g., Bobko & Karren, 1979; Cleveland & McGill, 1984). Commonly, researchers attempt to model this using either Weber or Fechner’s laws. Importantly, some research suggests that large correlation values are also underestimated while some research suggests that individuals struggle to distinguish meaningful estimates from correlations lower than .2 (Doherty et al., 2007). The body of research suggests that there is a systematic perceptual factor that facilitates individuals to over-or-underestimate correlations when examining scatterplots in a variety of forms, e.g., varying opacity, colour, and size. Thus, designing scatterplots more accustomed to accurate estimations should attempt to rectify this.

A final point regarding perceptual drivers is one of visual factors. (Wang et al., 2022) furthered the proposition that there are three levels of visualisation understanding: elementary level, intermediate level, and advanced level. The elementary level refers to reading specific values on the graph, e.g., how many data points are on the scatterplot (Garfield & Ooms, 2005). The intermediate level refers to understanding trends or relationships on the graph, e.g., what level of correlation is on the scatterplot (Carswell, 1992). The advanced level refers to understanding beyond the graph, e.g., this graph shows a correlation of .7 which differs from a prior graph assessing the same relationship (Börner et al., 2019; Borner & Maltese, 2015; Boy et al., 2014). Typically, correlation perception is considered solely with the elementary and intermediate levels. Thus, correlation perception the ability of an individual to assess elementary and intermediate visual understanding. Overall, drivers of correlation perception is based on cognitive abilities, such as using mean distances within the scatterplot and having intermediate visualisation understanding and a systematic estimation bias amongst individuals (Wang et al., 2022).

**Methods of modifying scatterplot perception**

To improve scatterplots so that people are more able to estimate correlation values, modifying the features of scatterplots presents an important method. The most basic scatterplot can be seen as having uniform dot size that is black-coloured and minimal or no labelling, such as x and y labels, scales, title (Rensink & Baldridge, 2010; Strain et al., 2023). From this form which we will denote as *S*, a scatterplot can then be manipulated in various ways and rigorously tested to assess whether the modifications have produced a change in correlation perception. With minimal changes, e.g., 1-2, researchers can then assess whether the modifications to *S* are beneficial or redundant. However, the changes should ideally not violate Tufte and Hehman and Xie (2021) laws. For example, within scatterplots, researchers could remove the plot points and simply show the regression line that is a function of the r value; this may or may not lead to improved estimation. However, this violates Tufte’s rule of showing as much data as possible. Further, researchers could colour code r values; for example, graphs could be colour coded as green = .7+, yellow = .4-.69, red = <.4. However, this could lead to scatterplots being non-inclusive to colourblind individuals and conflate estimation with signposted guesswork (Hehman & Xie, 2021).

There are numerous ways to modify *S*, Sarikaya and Gleicher (2018) suggested that the modifications align across four categories: point encoding, point grouping, point position, and graph amenities. First, graph amenities relate to parts of scatterplots unrelated to the scatterplot dots: examples include axis titles, graph lines, and scale labels. Second, point position regards modifying the position of the points to either emphasise certain parts, e.g., zooming or displacing, or adding information, e.g., subsampling and animation. Third, point grouping refers to the changing of scatterplot dot groupings; for example, shape abstraction, e.g., converting the dots to a shape, or shape enclosure, e.g., adding a regression line or a density plot. Last, the most common type of modification is point encoding whereby the points are modified, e.g., change of size, shape, colour, or opacity. The key feature of extensive research is point encoding whereby the features of the scatterplot points are manipulated in various ways to attempt to achieve changes in correlation perception (Ip et al., 2021; Strain et al., 2023).

**ADD REFERENCES FROM SARIKAYA**

**Size**

The most common and sound way of modifying a scatterplot is point encoding. Further, as the focus of this paper is modifying the size of points, it is important to understand what research has been conducted on this type of point encoding. For example, Rensink (2012) found that manipulating the size of scatterplot dots, e.g., making them smaller or larger, did not influence correlation perception. However, this was a relatively small study with only 20 participants per condition. Further, Micallef et al. (2017) developed a novel loss function method and manipulated the size and opacity of points on scatterplots. While changes to size significantly influenced outlier detection and class detection, there was no significant differences regarding correlation estimation (Micallef et al., 2017).

Further, Ip et al. (2021), building upon the work of Rensink (2012), compared five dot point conditions of 1 mm, 3 mm, 5 mm, 8 mm, and a mixed condition of the four prior conditions against a plot of 65 mm by 65 mm. While they found no differences in correlation estimation based on size, the sample was small (n = 18). Importantly, under traditional manipulations, size appears to not influence correlation estimation; however, size does appear to influence perceptions in two ways. First, larger points are more salient (Healey & Enns, 2012). Participants notice larger points more readily and easily than smaller points (Healey & Enns, 2012). Second, with these larger points, participants are biased towards these (Hong et al., 2022). For instance, when assessing descriptive statistics from scatterplots, participants will judge the mean based on larger points rather than smaller points (Hong et al., 2022). Thus, there is the potential for methods to be used to take advantage of this bias. A novel methodology has recently been developed that could prove beneficial for size as an effective point encoding method.

Smart (conference paper) Christian van Onzenoodt

**Strain et al’s (2023) non-linear decay transformation**

Recent developments by Strain et al. (2023) have demonstrated that transforming point encoding factors using a non-linear transformation have enhanced correlation perceptual estimation. Strain et al. (2023) modified contrast, also known as opacity, in a repeated measures study with 150 participants using four conditions: a full contrast condition, a linear condition, a nonlinear decay condition, and an inverted nonlinear decay condition. The results demonstrated that the nonlinear decay condition facilitated participants to estimate correlations significantly better than each of the other conditions. The nonlinear decay transformation is described below.

Within the transformation, 0.25 was selected as the value of b due to prior literature reporting underestimation. Further, r refers to the residual within the plot. Thus, the alpha or contrast level was calculated via this function. Points further away from a residual value of zero had lowered opacity. Importantly, this nonlinear decay transformation has implications for other point encoding features. If this method can improve correlation perception when using contrast, it is important to see if this can improve correlation perception when using the point encoding feature of size.

**Present study**

The present study will employ the methodology used by Strain et al. (2023) to examine whether the nonlinear decay transformation can improve correlation perception by modifying the size of the plot points. Strain et al. (2023) employed a robust and strenuously designed study to experimentally examine which condition was the most effective for correlation perception. For example, each participant (n=150) examined 180 plots (45 for each condition). Thus, the study had approximately 27,000 observations and each condition had approximately 7000 observations. Using Mayo's (2018) distinction, it passed strong severe testing. Strong severe testing is when a study is designed well enough to find discrepancies, errors, or counterevidence for *A* (Meehl, 1990). If few or none are present in the study data *X*, then there is evidence for *A* (Mayo, 2018). Thus, we will employ similar design features to severely test whether the nonlinear transformation is superior to other conditions. For example, we will employ a similar number of participants and stimuli, and similar experimental conditions and features, e.g., a fully reproducible study. Further, due to the study design, e.g., high number of observations, and the plots representing *S* (no titles, axes, ticks, etc), any differences in condition mean error scores will represent a difference in conditions, not some other variable. Overall, the study will have the following hypotheses:

H1: The full model including conditions will predict significantly more when compared with a null model without conditions.

H2: The nonlinear decay condition will be significantly lowest error rates among conditions.

H3: The inverted nonlinear decay will have the significantly highest error rates among conditions.

Exploratory model comparisons: akin to Strain et al. (2023), several models will be compared to the full model of H1 to assess whether additional factors are important. Four total model comparisons will be implemented: a literacy model, a training model, a dot pitch model, and an objective r category model. Each model will be compared to the basic model to assess whether there are significant differences and where these differences lie. Further, unlike some prior literature, each model will be built with an interaction term for the added variable.

**Method**

**Open research statement**

**Participants**

***Participant characteristics***

164 participants were recruited via Prolific, however, 14 participants failed attention checks within the experiment. Thus, the final sample had 150 participants (female = 75, male = 72, nonbinary = 3) with a mean age of 29.60 (*SD*= 8.54).

***Sampling procedure and inclusion/exclusion criteria***

Participants were required to have English language fluency and normal-to-corrected normal vision. Similar to Strain et al. (2023), data quality guidance from Peer et al. (2022) was implemented. Prolific users were required to have at least a 95% approval rating, have completed at least 100 prior submissions, and have not conducted studies previous studies by the authors, e.g., Strain et al. (2023). Additionally, when participants conducted the experiment, they had six attention checks, if they only two or less correct, they were removed from the study results.

***Sample size and statistical power***

The present study implemented similar sample size requirements to Strain et al. (2023) whereby approximately 150 participants were required. As the study was a repeated measures design, this meant each condition had 150 participants and there was a total of approximately 27,000 observations. Further, in study 2, Strain et al. (2023) highlighted that the sample size was powerful enough to detect significant results and small-to-medium effect sizes.

**Materials**

***Demographic and graph literacy questions***

Participants were asked for their age and gender.

***Subjective graph literacy scale*** (Garcia-Retamero et al., 2016)

Participants answered five questions related to their graph literacy (Garcia-Retamero et al., 2016), e.g., “How good are you at working with bar charts?”. Literacy questions were rated on a six-point scale ranging from *not at all* (0) to *extremely well* (5). The scale boasted excellent internal consistency, as assessed by Cronbach’s alpha using the lmt package, a = 0.914, Bootstrap (1000 samples) 95% confidence intervals [0.887, 0.936].

***Plot generation***

The experimental plots were generated in accordance with the methodology used by Strain et al. (2023). Each plot was generated as a 1000 x 1000 pixel .png image. Each plot had 128 plot points and included no titles, subtitles, scales, or ticks. Overall, there were 45 different r value scatterplots and each r value was shown four times – one for each condition. The r values ranged from .2 to .99 as prior research has suggested that little-to-no correlation is found at and below .2 r.

***Experimental measures***

Each generated plot had an objective r value, e.g., .7. Participants were asked to estimate said r value, e.g., a participant estimated .6 on a slider of 0 to 1. A response difference variable was calculated via objective r value – participant r value. These response difference values were then transformed into an absolute number (all positive numbers). This allowed for accurate comparisons between conditions; this is because an overestimation would lead to a negative number, whereas an underestimation would lead to a positive number. It is therefore possible for a condition that performs poorly to have high error rates at both the positive and negative range. This could lead to a low overall error rate because of the averaging of non-absolute numbers. Researchers then may incorrectly suppose that the condition performs well. Thus, making the values absolute numbers allowed us to compare differences in r across conditions regardless of under-or-overestimation and minimises issues due to high variances of correlation estimation.

***Dot pitch***

Prior studies, e.g., Strain et al. (2023), could not infer dot pitch from the experiment. However, this study employed the ScreenScale methodology (Morys-Carter, 2021) whereby participants are asked to modify a credit card on the screen to be the size of a credit or debit card they own. Credit, debit, and ID cards are a universal size of 85.6mm x 53.98mm (width x height). Within the psychopy experiment, ScreenScale infers screen height in cm. Further, psychopy infers pixel width and height. Thus, we were able to calculate dot pitch using the below formula.

Within the study, we assumed a 16: 9 (width: height) aspect ratio which is a ratio of 0.5625 whereby height is 0.5625 times the size of width and width is 1.77778 times the size of height. The pixel width was multiplied by 0.5625 to get pixel height. Further, screen height was multiplied by 1.77778 to get width and then divided by 2.54 to convert it into inches. Finally, the formula was produced. The mean dot pitch was 0.354 mm (SD = 0.0623).

***Visual threshold testing***

Visual threshold testing was conducted to ascertain whether all plot points within the experiment were visible. It is essential that each plot point is visible considering the study examines correlation perception of scatterplots. This was examined by asking participants to state the number of plot points in six graphs shown. 142 participants scored 6 out of 6 and 8 participants scored 5 out of 6 (mean score = 5.95, sd = 0.225). Overall, these results suggest that the experimental design passed visual threshold testing and that each participant was able to see the full number of plot points.

**Design**

***Experiment***

The experiment was designed using psychopy and hosted on pavlovia. The experimental design was a repeated measures design with one IV and one DV. Each participant was in all conditions and saw all plots. The IV was condition and the DV was the absolute response difference of the objective r value minus the participant rated r value. Additionally, random effects were collected: participants and items. Finally, for additional model building, other IVs were collected, e.g., dot pitch, graph literacy, training, and demographic factors. Within the experiment, each plot was shown to participants in a random order.

***Conditions***

Overall, there were four conditions within the study. Within each condition, the minimum dot size was 12 pixels. A scaling factor of 4 and a constant of .2 was used to produce the size modification in the scatterplots. The standard condition had the minimum dot size for each point with no changes in scale. The linear condition had an equation whereby the residuals were divided by 3.2 and as the residuals got further away from the regression line, the points decreased in size.

Like Strain et al., (2023), the nonlinear decay transformation was used and described below. Within the transformation, 0.25 was selected as the value of b due to it being used in the study by Strain et al. (2023). Further, r refers to the residual within the plot. Thus, the size level was calculated via this function. Points further away from a residual value of zero were nonlinearly smaller.

The final condition was the inverted nonlinear decay condition. The equation is described below. As the residuals got further away from the regression line, they nonlinearly got larger.

**A collage of black dots

Description automatically generated**

**Procedure**

***General procedure***

Once participants were recruited via Prolific, they started the experiment at their convenience. At the start of the experiment, participants were shown a participant information sheet and then provided consent to participate in the study using keyboard key presses. Participants then provided their age and gender identity and completed the subjective graph literacy test and visual threshold tests. Prior to engaging with the stimuli, participants completed the screenScale test. Participant were then shown instructions and examples of r levels, e.g., .2, .5, .8, and .95. Participants then were shown two practice trials before working through the series of 180 randomly shown plots. Before these plots, a visual mask was shown to participants. Participants used a slider ranging from 0 to 1 to estimate the r value shown in the scatterplot. Throughout the trials, six attention checks were displayed whereby participants were asked to either put the slider to 0 or 1 or ignore the plot. After completing the trials, participants were paid and those who completed two or more attention checks data were exported. Data was then statistically analysed in R.

***Data diagnostics***

The predominant method for ensuring the analyses used are relevant to the data will be using assumption tests from the performance package. This will assess things such as homogeneity of variance, normality of residuals, and outliers.

***Analytic strategy***

To test the hypotheses, several statistical analyses will be conducted using the R programming language. To test H1, a linear mixed effect model will be built and tested against a null model. To test H2 and H3, pairwise comparisons will be conducted if H1 is supported. When conducting exploratory model comparisons, models with additional components will be tested against the H1 model via likelihood ratio testing to assess if there is a significant difference. If any of these models are significantly better than the H1 model, pairwise comparisons will be conducted.

To further strengthen the findings of any pairwise comparisons, standardised and unstandardised effect sizes will be reported to provide insight into the magnitude of the effects found (Cohen, 1990; Kelley & Preacher, 2012). For instance, the standardised effect size of Cohen’s d will be reported. Further, Cohen’s (1990) labels will be used but will be discussed in the discussion section *effect sizes* as some argue that these can be misleading (Baguley, 2004, 2009). Further, error rate differences between groups will be reported as a proxy for an unstandardised effect size. Moreover, if pairwise comparisons are conducted, family-wise error protection, e.g., Tukey, will be implemented to lower the risk of type I errors.

**Results**

**Participant Flow**

**Individual differences**

***Age and gender***

***Correlation estimates***

As prior literature has suggested systematic under-or-overestimations in correlation perception, we assessed participants correlation estimates by condition and strength of correlation. This was calculated using the non-absolute measure of the r values whereby objective r – participant estimate. Means and confidence intervals for all condition estimates were positive meaning that across conditions participants systematically underestimated the r levels. Importantly, correlation estimate means in the weak correlation category (.2 to .4) were negative whereas they were positive in the moderate and strong categories. This means that participants systematically overestimated correlations at weak levels whereas they underestimated correlations at moderate to strong levels. Thus, this study did support some prior findings of a systematic underestimation overall. However, ADDDDDDD (See Table 1).

condition variable n min max median iqr mean sd se ci

*<fct>* *<fct>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>*

1 inverted un\_response 6687 -0.715 0.99 0.131 0.241 0.14 0.202 0.002 0.005

2 linear un\_response 6839 -0.725 0.989 0.025 0.212 0.041 0.198 0.002 0.005

3 nonlinear un\_response 6857 -0.796 0.995 0.009 0.202 0.026 0.19 0.002 0.005

4 standard un\_response 6838 -0.776 1 0.147 0.261 0.164 0.208 0.003 0.005

r\_cat variable n min max median iqr mean sd se ci

*<fct>* *<fct>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>*

1 weak un\_response 6542 -0.796 0.38 0.023 0.277 -0.006 0.194 0.002 0.005

2 moderate un\_response 6536 -0.603 0.574 0.11 0.3 0.101 0.201 0.002 0.005

3 strong un\_response 14143 -0.404 1 0.082 0.234 0.134 0.203 0.002 0.003

**Hypothesis 1**

A linear mixed model was built to determine whether a model containing the IV of size would be significantly better at explaining the DV than a null model without size. The buildmer package was used to assess the most complex stable model from the most complex model of response ~ condition + (1 + condition | item) + (1 + condition | participant). Overall, the final model that fit the data including condition and random intercepts for participant and item. The performance package revealed that the model met the assumptions (See figure x).

A likelihood ratio test was conducted; the full model was significantly better at explaining the DV than the null model, X2 (3) = 1373.4, p < .001. Thus, the effect of size was deemed to be significant. Figure X displays the full model’s conditions in violin-sina-boxplots or sitar plots (for simplicity and expediency) using ggplot2. A guitar plot contains a violin plot that measures population density of the sample, a sina plot that demonstrates this density via dot plots to show outliers, and a box plot to display the median and quartile ranges.

**A graph of a graph of a graph

Description automatically generated with medium confidence**

**Hypothesis 2 and 3**

To examine hypotheses 2 and 3, Tukey-corrected posthoc tests were conducted on the final model using the emmeans package. The estimated marginal means (see table X) suggested that nonlinear decay had the lowest error rate, followed by the linear condition, then the inverted condition, and finally the standard condition. Pairwise comparison tests demonstrated that each comparison was significant. See table X for all posthoc tests, z-scores, p-values, and cohen’s d effect sizes all computed using the emmeans package.

|  |  |  |  |
| --- | --- | --- | --- |
| Condition | Estimated marginal mean | Lower confidence level | Upper confidence level |
| Nonlinear | 0.138 | 0.124 | 0.152 |
| Linear | 0.147 | 0.133 | 0.162 |
| Inverted | 0.193 | 0.179 | 0.208 |
| Standard | 0.205 | 0.191 | 0.219 |

condition emmean SE df asymp.LCL asymp.UCL

inverted 0.193 0.0073 Inf 0.179 0.208

linear 0.147 0.0073 Inf 0.133 0.162

nonlinear 0.138 0.0073 Inf 0.124 0.152

standard 0.205 0.0073 Inf 0.191 0.219

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Contrast | Estimated difference | Z ratio | P value | Cohen’s d | Cohen’s d lower confidence level | Cohen’s d upper confidence level |
| Nonlinear – linear | -0.00995 | -4.357 | .0001 | 0.0743 | 0.0409 | 0.1078 |
| Nonlinear – standard | -0.06676 | -30.396 | <.0001 | 0.5194 | 0.4856 | 0.5532 |
| Nonlinear – inverted | -0.05529 | -25.064 | <.0001 | 0.4302 | 0.3963 | 0.4640 |
| Linear – inverted | -0.4574 | -20.730 | <.0001 | 0.3558 | 0.3221 | 0.3896 |
| Linear – standard | -0.5721 | -26.041 | <.0001 | 0.4451 | 0.4114 | 0.4788 |
| Inverted – standard | -0.01147 | -5.197 | <.0001 | 0.0892 | 0.0556 | 0.1229 |

contrast estimate SE df z.ratio p.value

inverted - linear 0.04574 0.00221 Inf 20.730 <.0001

inverted - non\_linear 0.05529 0.00221 Inf 25.064 <.0001

inverted - standard -0.01147 0.00221 Inf -5.197 <.0001

linear - non\_linear 0.00955 0.00219 Inf 4.357 0.0001

linear - standard -0.05721 0.00220 Inf -26.041 <.0001

non\_linear - standard -0.06676 0.00220 Inf -30.396 <.0001

contrast effect.size SE df lower.CL upper.CL

(inverted - linear) 0.3558 0.0172 191219 0.3221 0.3896

(inverted - non\_linear) 0.4302 0.0173 196555 0.3963 0.4640

(inverted - standard) -0.0892 0.0172 196941 -0.1229 -0.0556

(linear - non\_linear) 0.0743 0.0171 26479 0.0409 0.1078

(linear - standard) -0.4451 0.0172 26706 -0.4788 -0.4114

(non\_linear - standard) -0.5194 0.0172 27349 -0.5532 -0.4856

Further, hypothesis 2, that the nonlinear decay function will have the significantly lowest error rate was supported. The nonlinear condition had significantly lower mean error rates than all other conditions. Importantly, it also boasted medium and small-to-medium effect sizes when compared against the standard and inverted conditions, respectively. Further, it also demonstrated a very small effect size when compared with the linear condition.

However, hypothesis 3, that the inverted nonlinear decay condition will have the highest error rate was not supported. While it had significantly higher error rates than the nonlinear and linear conditions, it had a lower error rate than the standard condition. The comparisons between the inverted condition and the nonlinear and linear conditions boasted a small effect size whereas the comparions between the inverted and standard condition boasted a very small effect size.

**Additional Models**

Finally, in convention with strain et al. (2023), it is interesting to conduct exploratory model building to assess whether various other variables can improve on the final model performance. For instance, Strain et al. (2023) assessed whether training effects (first half of experiment vs second half of experiment) influenced the model. While they found no model improvements, it is worthwhile to assess whether similar factors can improve the current model’s performance.

**Literacy**

[**https://stats.stackexchange.com/questions/598594/how-do-you-conduct-contrasts-pairwise-comparisons-for-a-lmer-when-your-iv-is**](https://stats.stackexchange.com/questions/598594/how-do-you-conduct-contrasts-pairwise-comparisons-for-a-lmer-when-your-iv-is) **IF i want to add more about this.**

[**https://cran.r-project.org/web/packages/emmeans/vignettes/interactions.html**](https://cran.r-project.org/web/packages/emmeans/vignettes/interactions.html)

First, like Strain et al. (2023), a model with the addition of graph literacy was built and compared against the size model. Unlike Strain, however, who produced a graph literacy model with an additive term using buildmer’s add.terms function, e.g., condition + literacy, we implemented an interaction model: condition\*literacy. The literacy model explained significantly more variance than the size model, X2 (4) = 35.626, p < .001. The emmeans package’s emtrends function was used to conduct pairwise comparisons. The emtrends function allows for a categorical predictor to be paired with a continuous predictor when conducting pairwise comparisons. Figure X reveals that as graph literacy increases, the conditions each reduce in error. However, as the figure demonstrates, there are clear differences between conditions at lower-to-medium graph literacy levels. Further, the interaction appears to be driven by two non-significant findings when comparing the fitted slopes of each condition. First, non-linear, while boasting a lower mean error rate, was not significantly different when compared to linear; this because of the mean similarity at higher graph literacy levels. Further, the linear condition was not significantly different to the inverted condition; this was because the unit increase rate of graph literacy was not significantly different. Lastly, while there are two non-significant findings, all the other comparisons remain significant. Table X displays the estimated marginal means of condition based on a unit increase of graph literacy. Table X shows the pairwise comparisons between conditions when graph literacy is accounted for. Finally Figure Y shows a line plot of error rate by graph literacy whereby the lines are grouped by condition.

|  |  |  |  |
| --- | --- | --- | --- |
| Condition \* Graph literacy | Estimated condition trend based on 1 unit increase of graph literacy | Lower confidence level | Upper confidence level |
| Nonlinear | -0.000151 | -0.00290 | -0.002601 |
| Linear | -0.000929 | -0.00368 | -0.001824 |
| Inverted | -0.001687 | -0.00444 | -0.001067 |
| Standard | -0.003069 | -0.00582 | -0.000317 |

condition sum\_lit.trend SE df asymp.LCL asymp.UCL

inverted -0.001687 0.00141 Inf -0.00444 0.001067

linear -0.000929 0.00140 Inf -0.00368 0.001824

nonlinear -0.000151 0.00140 Inf -0.00290 0.002601

standard -0.003069 0.00140 Inf -0.00582 -0.000317

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Contrast | Estimate | Z ratio | P value | Cohen’s d | Cohen’s d LCL | Cohen’s d UCL |
| Nonlinear – linear | 0.000778 | 1.516 | 0.4281 | 0.00606 | 0.00178 | 0.01389 |
| Nonlinear – inverted | 0.001538 | 2.979 | 0.0153 | 0.01196 | 0.00409 | 0.01983 |
| Nonlinear – standard | 0.002919 | 5.687 | <.0001 | 0.02273 | 0.01490 | 0.03057 |
| Linear – inverted | 0.00758 | 1.469 | 0.4565 | 0.00591 | 0.00198 | 0.01379 |
| Linear – standard | 0.002141 | 4.167 | 0.0002 | 0.01668 | 0.00883 | 0.02452 |
| Inverted - standard | 0.001383 | 2.678 | 0.0372 | 0.01077 | 0.00289 | 0.01865 |

contrast estimate SE df z.ratio p.value

inverted - linear -0.000758 0.000516 Inf -1.469 0.4565

inverted - nonlinear -0.001536 0.000516 Inf -2.979 0.0153

inverted - standard 0.001383 0.000516 Inf 2.678 0.0372

linear - nonlinear -0.000778 0.000513 Inf -1.516 0.4281

linear - standard 0.002141 0.000514 Inf 4.167 0.0002

nonlinear - standard 0.002919 0.000513 Inf 5.687 <.0001

contrast effect.size SE df asymp.LCL asymp.UCL

(inverted - linear) -0.00591 0.00402 Inf -0.01379 0.00198

(inverted - nonlinear) -0.01196 0.00402 Inf -0.01983 -0.00409

(inverted - standard) 0.01077 0.00402 Inf 0.00289 0.01865

(linear - nonlinear) -0.00606 0.00400 Inf -0.01389 0.00178

(linear - standard) 0.01668 0.00400 Inf 0.00883 0.02452

(nonlinear - standard) 0.02273 0.00400 Inf 0.01490 0.03057

NOTE: WILL IMPROVE THIS GRAPH

**A graph on a black background

Description automatically generated**

**Training**

Similar to Strain et al. (2023), we built a model whereby the variable of training was implemented. This meant that the number of trials were split in half to assess whether training effects of the practice and example trials influenced correlation perception. Unlike Strain et al. (2023) an interaction model was built, condition \* training. This model did not explain significantly more variance than the size model, X2 (4) = 3.4896, p = .4795.

**Dot pitch**

As this study was able to measure dot pitch, it is interesting to examine whether this variable influenced response estimates. A model containing the interactive dot pitch variable was compared to the size model. The dot pitch model explained significantly more variance than the size model, X2 (4) = 34.183, p < .001. The interaction between dot pitch and condition was driven by the linear, inverted, and standard conditions being non-significant from each other. For these conditions, as dot pitch increased, mean error reduced. However, for the nonlinear condition, which boasted significantly lower error rates than all other conditions, as dot pitch increased mean error increased. When producing the effect sizes, a correction was conducted due to abnormally large effect sizes being produced, e.g., 1.9 UCL. The explanation and these non-corrected effect sizes are reported in Appendix X.

|  |  |  |  |
| --- | --- | --- | --- |
| Condition \* dot pitch | Estimated condition trend based on 1 unit increase of dot pitch | Lower confidence level | Upper confidence level |
| Nonlinear | 0.0577 | -0.132 | 0.2479 |
| Linear | -0.0550 | -0.245 | 0.1352 |
| Inverted | -0.1317 | -0.322 | 0.0568 |
| Standard | -0.1122 | -0.301 | 0.0790 |

condition dot\_pitch.trend SE df asymp.LCL asymp.UCL

inverted -0.1317 0.0971 Inf -0.322 0.0586

linear -0.0550 0.0971 Inf -0.245 0.1352

nonlinear 0.0577 0.0970 Inf -0.132 0.2479

standard -0.1112 0.0970 Inf -0.301 0.0790

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Contrast | Estimate | Z ratio | P value | Cohen’s d | Cohen’s d LCL | Cohen’s d UCL |
| Nonlinear – linear | 0.1128 | 3.184 | 0.0079 | 0.2365 | 0.0909 | 0.3821 |
| Nonlinear – inverted | 0.1894 | 5.325 | <.0001 | 0.3972 | 0.2510 | 0.5434 |
| Nonlinear – standard | 0.1690 | 4.775 | <.0001 | 0.3543 | 0.2089 | 0.4997 |
| Linear – inverted | 0.0776 | 2.152 | 0.1370 | 0.1607 | 0.0143 | 0.3070 |
| Linear – standard | 0.0562 | 1.586 | 0.3867 | 0.1778 | 0 = -0.0278 | 0.2634 |
| Standard – inverted | 0.0205 | 0.0575 | 0.9396 | 0.0429 | 0 = -0.189 | 0.1033 |

contrast estimate SE df z.ratio p.value

inverted - linear -0.0766 0.0356 Inf -2.152 0.1370

inverted - nonlinear -0.1894 0.0356 Inf -5.325 <.0001

inverted - standard -0.0205 0.0356 Inf -0.575 0.9396

linear - nonlinear -0.1128 0.0354 Inf -3.184 0.0079

linear - standard 0.0562 0.0354 Inf 1.586 0.3867

nonlinear - standard 0.1690 0.0354 Inf 4.775 <.0001

contrast effect.size SE df asymp.LCL asymp.UCL

(inverted - linear) -0.1607 0.0747 Inf -0.3070 -0.0143

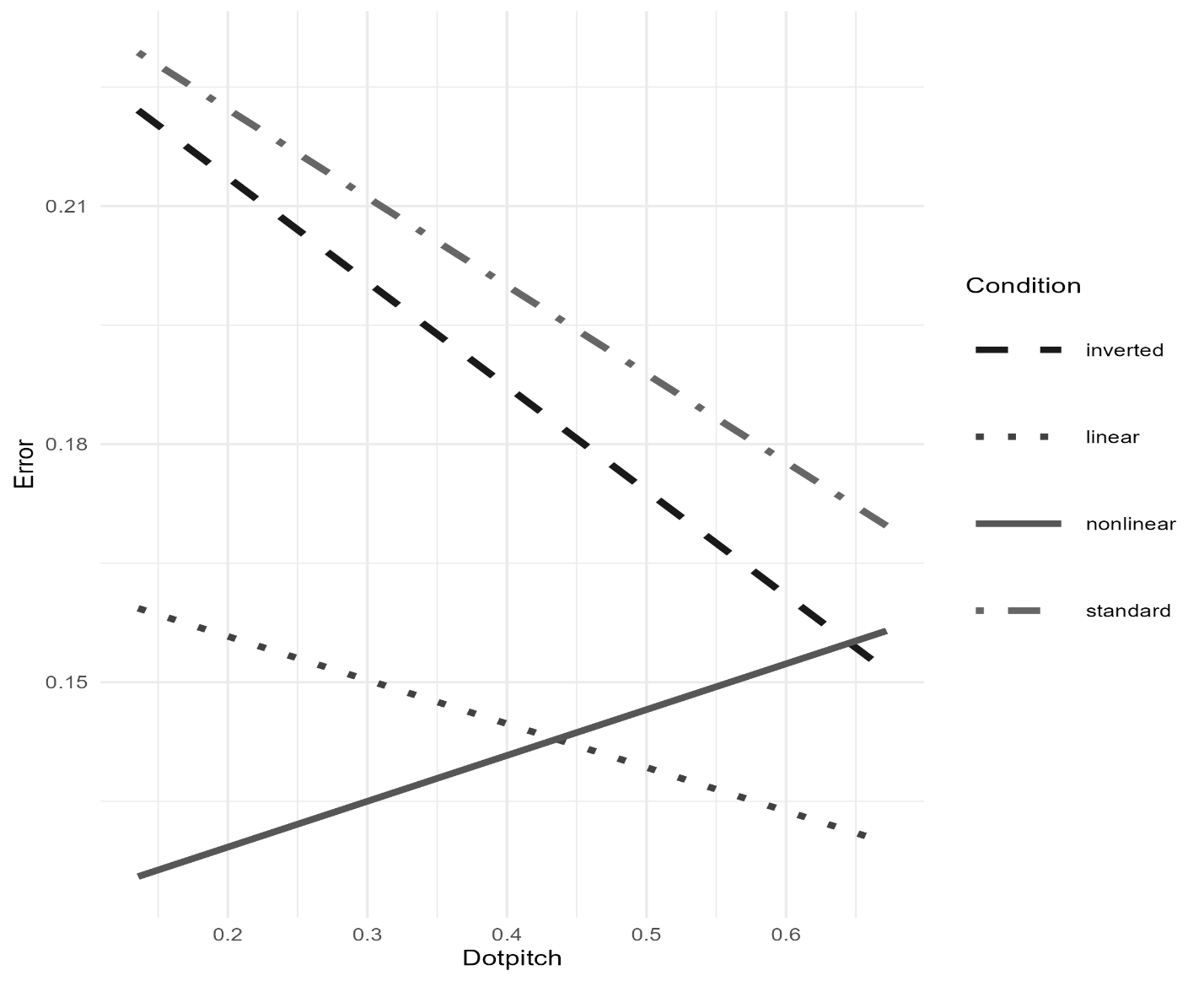
(inverted - nonlinear) -0.3972 0.0746 Inf -0.5434 -0.2510

(inverted - standard) -0.0429 0.0746 Inf -0.1891 0.1033

(linear - nonlinear) -0.2365 0.0743 Inf -0.3821 -0.0909

(linear - standard) 0.1178 0.0743 Inf -0.0278 0.2634

(nonlinear - standard) 0.3543 0.0742 Inf 0.2089 0.4997

****

**Levels of objective R**

A final model to assess the experiment is to include the objective r values as categories to assess whether the nonlinear decay function is consistent throughout the levels of objective r. The objective r values were transformed into categories where .2-.39 was weak, .4-.59 was moderate, and .6+ was strong. Sum contrast coding was set on condition and r value category using the stats package. This is because the mixed model used is implemented as an improvement of a factorial ANOVA. To correctly interpret a mixed model being used for this purpose such as examining main effects and interactions, sum coding should be used (Brehm & Alday, 2022). Otherwise, it is not uncommon to misinterpret simple effects as main effects (Brehm & Alday, 2022; Schad et al., 2020).

After contrasts were set, an interaction model of condition \* r category was built to assess whether the inclusion of r category would explain significantly more variance than the size model. Indeed, the interaction model did predict significantly more variance, X2 (8) = 855.36, p < .001.

Corrected pairwise comparisons were conducted to assess whether there were significant differences in condition responses at the same objective r category, e.g., linear weak vs standard weak. The results suggested that the interaction was being driven at the moderate and strong correlation levels whereby conditions significantly differed. At the moderate level, all comparisons were significant except for the linear vs non-linear comparison, p = 1.00. Further, at the strong level, all comparisons were significant except for the standard vs inverted comparison, p = 1.00. Figure X displays a sitar plot of error rate by condition that is facetted by objective r strength category.

A screenshot of a graph

Description automatically generated

linear moderate - non\_linear moderate 0.00658 0.00441 Inf 1.492 1.0000

inverted strong - standard strong -0.00598 0.00303 Inf -1.975 1.0000

**Interestingly, at the weak level, all comparisons were insignificant.**

**Estimated marginal means table**

condition r\_cat emmean SE df asymp.LCL asymp.UCL

inverted weak 0.2154 0.0113 Inf 0.1932 0.238

linear weak 0.2297 0.0113 Inf 0.2075 0.252

non\_linear weak 0.2205 0.0113 Inf 0.1982 0.243

standard weak 0.2197 0.0113 Inf 0.1975 0.242

inverted moderate 0.1914 0.0157 Inf 0.1607 0.222

linear moderate 0.1674 0.0157 Inf 0.1366 0.198

non\_linear moderate 0.1608 0.0157 Inf 0.1300 0.192

standard moderate 0.2264 0.0157 Inf 0.1956 0.257

inverted strong 0.1835 0.0104 Inf 0.1631 0.204

linear strong 0.1016 0.0104 Inf 0.0812 0.122

non\_linear strong 0.0906 0.0104 Inf 0.0702 0.111

standard strong 0.1895 0.0104 Inf 0.1692 0.210

**CONTRASTS**

contrast estimate SE df z.ratio p.value

inverted weak - linear weak -0.014093 0.00442 Inf -3.189 0.0942

inverted weak - nonlinear weak -0.005338 0.00442 Inf -1.208 1.0000

inverted weak - standard weak -0.004739 0.00442 Inf -1.072 1.0000

linear weak - nonlinear weak 0.008755 0.00442 Inf 1.982 1.0000

linear weak - standard weak 0.009354 0.00442 Inf 2.116 1.0000

nonlinear weak - standard weak 0.000599 0.00442 Inf 0.136 1.0000

inverted moderate - linear moderate 0.024373 0.00442 Inf 5.509 <.0001

inverted moderate - nonlinear moderate 0.030655 0.00442 Inf 6.935 <.0001

inverted moderate - standard moderate -0.035064 0.00442 Inf -7.925 <.0001

linear moderate - nonlinear moderate 0.006282 0.00442 Inf 1.422 1.0000

linear moderate - standard moderate -0.059437 0.00442 Inf -13.438 <.0001

nonlinear moderate - standard moderate -0.065718 0.00442 Inf -14.872 <.0001

inverted strong - linear strong 0.082385 0.00304 Inf 27.144 <.0001

inverted strong - nonlinear strong 0.093129 0.00303 Inf 30.712 <.0001

inverted strong - standard strong -0.006134 0.00304 Inf -2.019 1.0000

linear strong - nonlinear strong 0.010744 0.00299 Inf 3.594 0.0215

linear strong - standard strong -0.088519 0.00300 Inf -29.482 <.0001

nonlinear strong - standard strong -0.099263 0.00300 Inf -33.088 <.0001

**EFFECT SIZE**

contrast effect.size SE df asymp.LCL asymp.UCL

(inverted weak - linear weak) -0.11151 0.0350 Inf -0.180049 -0.04297

(inverted weak - nonlinear weak) -0.04224 0.0350 Inf -0.110757 0.02628

(inverted weak - standard weak) -0.03749 0.0350 Inf -0.106057 0.03107

(linear weak - nonlinear weak) 0.06928 0.0350 Inf 0.000753 0.13780

(linear weak - standard weak) 0.07402 0.0350 Inf 0.005452 0.14258

(nonlinear weak - standard weak) 0.00474 0.0350 Inf -0.063810 0.07329

(inverted moderate - linear moderate) 0.19285 0.0350 Inf 0.124220 0.26149

(inverted moderate - nonlinear moderate) 0.24256 0.0350 Inf 0.173977 0.31114

(inverted moderate - standard moderate) -0.27744 0.0350 Inf -0.346098 -0.20879

(linear moderate - nonlinear moderate) 0.04970 0.0350 Inf -0.018828 0.11824

(linear moderate - standard moderate) -0.47030 0.0351 Inf -0.539006 -0.40159

(nonlinear moderate - standard moderate) -0.52000 0.0350 Inf -0.588673 -0.45133

(inverted strong - linear strong) 0.65188 0.0242 Inf 0.604491 0.69926

(inverted strong - nonlinear strong) 0.73689 0.0242 Inf 0.689459 0.78432

(inverted strong - standard strong) -0.04854 0.0240 Inf -0.095651 -0.00143

(linear strong - nonlinear strong) 0.08501 0.0237 Inf 0.038652 0.13137

(linear strong - standard strong) -0.70042 0.0239 Inf -0.747351 -0.65348

(nonlinear strong - standard strong) -0.78543 0.0240 Inf -0.832421 -0.73844

**MUST DO THIS SOON**

**Discussion**

**Hypotheses**

Overall, the study adds to prior literature via supporting and rejecting several hypotheses; specifically, First, H1 was supported. Size as a fixed effect explained significantly more variance than a null model without size. Thus, this suggests that the experiment was able to provide evidence that the manipulations used in the study did influence correlation estimate error rates. Second, H2 was supported, we found that the nonlinear decay transformation provided the lowest error rates among conditions. This suggests that the nonlinear decay condition facilitated participants to provide more accurate estimations of correlations when compared to the other conditions. Finally, H3 was not supported. The inverted nonlinear condition did not have the highest error rates - the standard dot plot size condition did.

With regard to the mentioned hypotheses, results can be compared to Strain et al. (2023). This is because we implemented similar methodologies and hypotheses to Strain. Like Strain, the analyses supported H1 and H2. Thus, the claim that the nonlinear transformation improves correlation estimation is further strengthened. However, we did not find support for H3, whereas Strain found evidence that the inverted nonlinear decay condition was the worst performing condition. This discrepancy could be explained by the use of a different point encoding modification: size. From the results of our study, it can be supposed that when compared to the standard dot size, variations of nonlinear and linear transformations to the size of the dots improve correlation perception. On the other hand, reducing opacity closer to the regression line appears to reduce performance.

ADD MORE LITERATURE HERE……

EXPLORATORY ANALYSIS JUSTIFICATION

Additionally, like Strain, we performed exploratory model comparisons. While these model comparisons can provide interesting and potentially useful insights, caution must be applied. The results from these exploratory analyses are simply that; they are not supported predictions or a priori (preregistered) hypotheses. While the model comparisons have an inherent hypothesis structure where the alternative hypothesis assumes one model is significantly better than the other, there were no a priori claims of direction. Some such as Nosek…. Would argue that these explorative analyses are tentative or less certain than a preregistered hypothesis. However, others like Rubin and Donkin (2022) argue that this entirely depends upon the type of exploratory analysis and the contextual factors surrounding the aforementioned test. Further, Szollosi et al. (2020) makes the argument that good research is good research regardless of preregistered analyses. Finally, using the same data for exploration and confirmation is perfectly fine (Pashler & Harris, 2021). The following section will provide two claims for why this study’s exploratory analyses are useful.

First, in accordance with numerous sources of open and transparent science, these analyses are fully reported, justified, disclosed, and reproducible. INSERT GITHUB…. No selective reporting has taken place and all tests and analysis code are provided. The results can be checked by other researchers who are able to make claims and interpretations regarding the research. Second, while some research has demonstrated that some analyses that are exploratory can be low in severity testing, the current study had severe testing. Ostensibly, each variable added to the model could reduce the strength of the study’s hypotheses, specifically H1 and H2.

Additionally, the exploratory analyses were not hypothesised after the results were known nor were hypotheses changed or retrieved, they were reported and interpreted (Rubin, 2017). Thus, they passed the falsification test outlined by preregistrationists (Rubin, 2022). Further, the testing followed acceptable and stringent rules. For example, type I error control. Most importantly, the results of these exploratory analyses are interpreted cautiously, the significance or non-significance of findings are not exaggerated or confirmatory. Thus, while some may argue these analyses are tentative, the philosophy of science arguments employed suggests that the analyses can be meaningfully interpreted.

Unlike Strain et al., (2023) interaction terms were added to all exploratory models. This was to examine whether condition interacted with these additional variables and if so, how they influenced each other. Graph literacy was added as an interaction term: we found a significant interaction between condition and graph literacy. While like Strain, graph literacy did not have a significant main effect, the interaction demonstrated that as graph literacy increases, each condition’s mean error reduced significantly bar the nonlinear decay condition. This may suggest that the nonlinear condition is beneficial for individuals of all graphical literacy levels whereas other conditions are primarily useful for individuals with higher graph literacy levels.

A commonly held view is that individuals with greater graph literacy are better at graphical interpretation. However, graph literacy as a main effect did not influence the differences between conditions, only the rate at which mean errors reduced or increased in conditions. Thus, it was concluded that while graph literacy interacted with conditions, it was not the main driver regarding correlation perception in participants, conditions were. This claim is further enhanced by the excellent internal consistency of the subjective graph literacy scale used to assess graph literacy. Thus, overall, while participant’s graph literacy influenced their ‘base’ level of error, the conditions they were in influenced their error rate more substantially.

TRAINING ANALYSES

Further, we added training level (first half, second half) as an interaction term and found that this model did not explain significantly more variance than the size model. This suggests that any potential priming effects from the example and training plots did not influence participant’s correlation estimates within the split we conducted. This split follows prior literature but it is possible that a different training split such as first, second, third, and fourth quarters could glean significance. Importantly, the training results also suggest that other order effects like fatigue, practice, and boredom effects are not significantly present within the first half second half training split we devised.

Unlike Strain et al., (2023), we implemented two additional exploratory analyses. First, dot pitch was added as an interaction term and added to an exploratory model. The model with dot pitch explained significantly more variance than the size model. The findings from the dot pitch model were atypical. A prior view regarding dot pitch might be that as the distance between pixels decreases and quality increases, the more easily an individual can view something. However, bar the nonlinear condition, the other three conditions had lower mean errors the greater the dot pitch. Thus, it appeared that in these three conditions, as quality decreased, mean error decreases also. Numerous conclusions can be drawn from this result for these three conditions. First, for these three conditions in this sample, it could be that higher dot pitch improves correlation estimation. Second, a more likely estimate is that, as dot pitch was a between-subjects effect, individual variance in performance meant that in this sample, the best performers also happened to have poorer quality monitors.

The nonlinear condition was influenced how traditional wisdom would assume dot pitch would work. This could also be explained by individual differences whereby the worst performers in the conditions were able to estimate correlation perception to a greater extent in the nonlinear condition than all other conditions whereas the effect was less pronounced with better performers or ones with higher dot pitch. Finally, the nonlinear condition boasts much lower variation in error than the other conditions; further, it was also significantly better than all other conditions.

Finally, an important addition to this work that was not conducted in prior exploratory analyses was the inclusion of correlation strength as a fixed effect in an interaction model. It was found that this inclusion explained significantly more variance than the size model. Further, it was found that at the weak correlation strength, .0-.39, little difference between conditions was noticed. However, at the moderate and strong correlation levels, most comparisons were significant. This is not entirely surprising as some prior literature has suggested that the weaker the correlation, the more difficult the interpretation; however, this was mainly regarding correlation strengths of .2 or less. Further, this finding suggests that much design care should be taken to improve weak correlation estimation. However, findings are limited to interpretation as the exploratory analysis was not a hypothesised result. Further, the categorising of correlations of weak, moderate, and strong neglect the subjectivity surrounding these categories. Future research assessing scatterplot design guidelines could design a more substantial variable to examine the effect of categorised correlation strength on estimates regarding other design variables.

Effect sizes

As mentioned in the method section, we will briefly discuss the usage of what Lenth (2001) has called ‘canned’ effect sizes. Within the results and parts of the discussion, we used Cohen’s (1990) language to interpret the effect sizes, e.g., .2 = small. While this is convention, it has been criticised as misleading by some researchers due to, among others, several issues. First, if one variable within a model of a specific sample is measured incorrectly or suffers from high variance, the effect size will be biased. Further, effect sizes in one study are difficult to compare to one in another study as researchers may use different sigma values (See Appendix A). Further, effect size interpretation are contingent on whether the effect is stable across measures; for instance, our effect sizes were larger than Strain et al. (2023). However, until we compare these within the same study, it is not known whether the nonlinear size modification is better than the nonlinear contrast modification. Further, a small effect size in one studied phenomena (medicine) can have far more practical significance than a large one in another phenomena (pharmacology; Schäfer & Schwarz, 2019). Finally, the use of ‘canned’ effect sizes (small, medium, large), specifically when assessing sample size, can be seen as a proxy for needing a small, medium, or large sample (Lenth, 2001). Thus, throughout the discussion, there has been a consistent call to the unstandardised effect size; for instance, the nonlinear condition had the lowest mean error rate.

IMPLICATIONS:

Implication 1: Systematic underestimation (mostly)

The first implication is that this study strengthens prior researchers claim that when people estimate correlation, they systematically underestimate. This finding was found throughout most of the sample. For instance, the mean estimation errors in each condition and at the moderate and strong correlation levels suggested that participants had systematically underestimated correlation strength. However, below r values of .39, the mean error suggested that participants overestimated correlation strengths. While this finding is interesting and goes against much literature suggesting a complete underestimation effect from r values of .2 +, it may not influence design parameters and suggestions as the next section will demonstrate. Simply, whether participants over-or-underestimate is less consequential than the fact that there are large errors in standard scatterplot designs.

Implication 2: Nonlinear decay works with size

The second implication is the study’s most promising finding; the nonlinear decay conditions demonstrate significant success when compared to the other conditions. In the base model, it was significantly better than all other conditions and supported the hypotheses entirely. It suggests that there is much promise in using this nonlinear method to improve average errors in correlation estimation. However, with this condition, there are some minor caveats. First, it appears that when correlation strength is considered, it is no better than the other conditions at catalysing participants to estimate correlations below .39; this is further discussed in implication 4. Further, when literacy is considered, higher graph literacy reduces the impact of the nonlinear function when compared with the linear condition. Simply, as the unit of graph literacy increases, the mean error difference between the linear and nonlinear conditions lessens. However, what should be noted is that this is because of the decrease in error of the linear condition, not error increases of the nonlinear condition. Overall, the nonlinear condition boasted favourable metrics and, within this study, is the superior condition.

Implication 3: modifying size improves estimation compared to standard plots

The third implication concerns hypothesis 3: the nonlinear decay would have the highest error rates among conditions. This hypothesis was not supported. Interestingly, the standard condition had the highest error rates among conditions. This finding suggests that when researchers are designing graphical displays, the point encoding feature of size should be considered. Specifically, researchers should manipulate the size of the points so that there is an obvious discrepancy – some small, some big. Ideally, researchers would implement the nonlinear decay feature, however, as was shown, all three conditions where superior to the standard conditions. Thus, the findings here demonstrate that modifying size improves estimation when compared to the standard condition.

Implication 4: smaller r values are more difficult to estimate

Finally, quite possible the most axiomatic but important implication regards small r values. Specifically, the findings of this research suggest that two axioms when r values are below .39. First, participants struggled to estimate r values at this level. This is consistent with prior research suggesting lower r values are estimated and discriminated poorly (Bobko & Karren, 1979; Cleveland & McGill, 1984; Doherty et al., 2007) Second, this effect was seemingly not improved by condition changes; for instance, in the r value category model, there were no significant differences among conditions. While this analysis was exploratory, it appears that conditions did not influence estimation at low r levels. It is therefore important for future research to consider how to improve upon this issue.

**Strengths**

The first strength relates to the study design; it followed suit with modern research and improved upon contemporary methods. For instance, Rensink had small samples and only used unitary manipulations such as large dots versus small plots. Here, the study was high powered and used theoretically sound manipulations that were aimed to increase or reduce estimation error.

The second strength relates to the study further supporting the nonlinear decay transformation. Here, the study has demonstrated that this manipulation can be used with more than one point encoding feature. Further, it shows that when this manipulation is used on the point encoding feature of size, it produces significantly better results than other manipulations.

A third strength of the study was that it abided by Gelman and Stewarts recommendations. All the data were analysed, no data was removed because it did not fit questionable criteria. Further, all comparisons were reported in easy-to-read table formats. In addition, measurements for variables were accurate and based on empirical research, e.g., the SGLS and estimation paradigms. An additional Gelman point was that the study was a fully repeated measures design which allows for higher powered studies.

A final strength related to Gelman and Stewarts recommendations is the promotion and use of open scientific practices. This includes making the analysis reproducible and the data and code accessible. Further, all analyses are fully reported and reproducible. Overall, this means that the public and other researchers can examine the script and check the claims made. This is an important component of good research as it improves transparency and lessens publication bias.

[**https://www.sciencedirect.com/science/article/pii/S0022103116301925?via%3Dihub#s9175**](https://www.sciencedirect.com/science/article/pii/S0022103116301925?via%3Dihub#s9175)

[**https://lakens.github.io/statistical\_inferences/13-prereg.html#how-to-preregister**](https://lakens.github.io/statistical_inferences/13-prereg.html#how-to-preregister)

**Cite seminar of Andrew** [**https://statmodeling.stat.columbia.edu/2013/05/17/how-can-statisticians-help-psychologists-do-their-research-better/**](https://statmodeling.stat.columbia.edu/2013/05/17/how-can-statisticians-help-psychologists-do-their-research-better/)

[**https://statmodeling.stat.columbia.edu/2017/06/05/advice-psychology-researchers-changed-since-2013/**](https://statmodeling.stat.columbia.edu/2017/06/05/advice-psychology-researchers-changed-since-2013/)

[**https://statmodeling.stat.columbia.edu/2016/08/24/balancing-bias-and-variance-in-the-design-of-behavioral-studies-the-importance-of-careful-measurement-in-randomized-experiments/**](https://statmodeling.stat.columbia.edu/2016/08/24/balancing-bias-and-variance-in-the-design-of-behavioral-studies-the-importance-of-careful-measurement-in-randomized-experiments/)

**LIMITATIONS:**

While the study possesses exciting implications for future research and rigorous strengths, it has certain limitations. This section will highlight five potential limitations of the study and how these could be mitigated and addressed in future research. The first limitation concerns the interpretation of the graph literacy model. While the scale boasted excellent internal consistency, time-efficiency, and has been used in several high-quality studies (e.g., Strain et al. (2023), it does not measure objective performance. It is possible that some participants who scored highly overestimated their abilities. If so, the graph literacy model can only be interpreted as a subjective variable. Specifically, the model units (Table X) could only be interpreted as when participant’s subjective view of their graph literacy increased, the mean estimation error reduced by X. Future research could employ objective measures of graph literacy, however, these may be time-consuming or provide order effects.

The second limitation concerns a theoretical issue regarding the beta value in the nonlinear decay function. Currently it is .25, however, this was developed by researchers providing best estimates for sensible values. It is possible that a different beta value, e.g., .26 or .27, could produce better performances of participants. While not an inherent limitation, the question of an optimal beta value is worth discussing if changes in beta improve or reduce correlation estimation performance.

Relatedly, a third limitation concerns the dot size manipulation. These were again decided by researchers providing best estimates. Within the size modified conditions, at the largest dot sizes, certain plots had dots that ‘merged’. As the goal of effective data visualisation is to make all plot points visible, the issue with dot size could have reduced estimative abilities of participants. This is because for certain plots, they would have observed dot clumps as opposed to singular dots. However, due to linear and nonlinear equations, this may be unavoidable due to constraints of screen and pixel sizes. Figure X demonstrates an example of this ‘merged’ clump of dots closer to the regression line.

A black dotted diagram with a white background

Description automatically generated

A potential fourth limitation regards the paradigm the study used: the direct estimation paradigm. This paradigm can be criticised in two ways. First, the concept of asking participants to estimate how large an r value is means that individual plots are not objectively measured, nor can they be compared to other singular plots; it is argued that this is because of the systematic bias in perceptual estimates and high variance in plot ratings. However, as the goal of the study was to reduce the systematic estimation bias, the criticism is dulled in two ways. First, while r value observations are not ‘objective’ (participant A’s .3 rating is not a universal .3 rating), the fact that mean errors were lowest in the nonlinear condition suggests that the manipulation reduced this systematic bias. Further, the sheer number of trials for each participant reduces the impact of high variance in the study.

The second estimation paradigm issue relates to the dependent variable. Accuracy was used to construct the dependent measure. While this is intuitive and allows us to ascertain over-and-underestimation, it does not allow us to infer complex relationships between the visualisation and correlation perception nor ascertain how a singular graph is understood when compared to other graphs. For instance, while the mean score for each condition was used to compare the condition, commentary on individual plot performance cannot be made due to the high variance in accuracy designs. Further, while objective r value plots were shown prior to the experimental stimuli, participants perceptions of what constitutes, say, a .7 correlation could widely vary. Further, while Strain et al. (2023) argues that participants make indirect comparisons, they are not true comparisons from plot to plot. Thus, like the direct estimation paradigm, the accuracy measure is limited by subjectivity of rating and high variance. It is argued that other methods such as the JND and discrimination paradigms are better equipped at dealing with ‘objective’ measures and providing explanations and context for singular plots. However, as the study was conducted to result in design implications, these methodologies may be less useful than direct estimation. This is because, in real world settings, people typically see plots in isolation rather than two plots together to see where the middle difference is.

A final potential limitation regards to the experimental design. First, seeing 180 sequential plots is not how people typically view visualisations in real-world settings. Further, in research and dissemination, plots are rarely shown as *S*, they typically have titles, colours, axes, and ticks. Thus, the findings may lack generalisability to real-world settings. However, this issue is mitigated by the findings of the study suggesting significant differences between conditions. Further, by using scatterplots *S* we were able to determine that the condition was the sole driver of changes in error and accuracy in the first model.

An additional component to this limitation is the possibility that the study design elicited extraneous effects, like fatigue effects, that may have biased study results and lowered or inflated performance. For instance, participants spent on average 39 minutes (SD = 14 mins) on the experiment whereby they went through 180 plots. While visual masks were present, no breaks or active rests, e.g., non-experimental tasks, were given to participants. Commonly, fatigue effects are present towards the end of the study; the exploratory analysis including the first and second half of the experiment suggested that fatigue, training, or practice effects were not significantly present. However, it is possible that these effects could have been present earlier than half-way through the experiment. Future research could further partition the timing to assess how fatigue or boredom influences this specific study design in exploratory analyses. However, much research suggests that fatigue effects are unavoidable regardless of countermeasure. Regardless of this, the findings were consistent with H1 and H2 and with prior findings of the same methodology.

**Future directions**

Overall, the study excelled and possessed several excellent strengths. Importantly, future research can build on this in several ways. First, the nonlinear decay transformation described and implemented in Strain et al. (2023) has been used with two-point encoding factors: opacity/contrast and size. Future research could add to this and implement additional point encoding features. For instance, a study could replicate the methodology used in this study but change the point encoding feature of size to other features such as colour, shape/symbol, texture/pattern, or boldness of outline among others. By examining additional point encoding features, the theoretical basis of the nonlinear decay transformation can be increased and eventually an ideal point encoding feature can be found.

Second, as the literature on negative correlations is lacking, researchers examining correlations and aiming to provide design implications should include negative correlation plots in their experiments. This would be beneficial for two reasons. First, this will falsify whether the nonlinear decay transformation improves correlation estimation in negative correlations. While one can assume that if it works with positive correlations it should work with negative ones, this is not known. Second, if the method works with negative correlation, it will strengthen design recommendations by having a technique that works across correlation strengths.

Third, while some may argue that the size and contrast experiments can be compared by the standardised effect size of Cohen’s d, others would disagree. Moreover, the high variance in correlation estimates means that this comparison is unideal. It cannot currently be said which point encoding feature works best. Thus, future research should compare point encoding features using the nonlinear decay transformation in the same experiment. This would be useful to assess whether, for instance, a size, contrast, or shape method is best individually.

Finally, the point encoding methods could be combined. For instance, a size and contrast experiment could be conducted comparing nonlinear size, nonlinear contrast, standard, and a nonlinear size-contrast conditions to assess if combining these transformations improve accuracy and reduce error size. This would be beneficial as there is potential to maximise the potency of the nonlinear decay transformation and produce data visualisations that people can estimate accurately and precisely.

**Conclusion**

The present study had the goal of assessing whether the nonlinear decay parameter described in Strain et al. (2023) would be effective when the point encoding feature of size was used instead of contrast. Importantly, H1 and H2 were supported. The nonlinear decay condition produced significantly lower error than the other conditions. However, the inverted condition did not possess the highest error rates – the standard size condition did. Further, additional analyses, while exploratory, suggested some important findings. First, graph literacy reducing the significant comparisons between standard-inverted and linear-nonlinear. Second, while the dot pitch model explained significantly more variance than the size model, the nonlinear condition remained the best condition. Finally, at the weak correlation level, .2 to .39, of the r category model participants error rates non-significantly differed between conditions. Overall, some important implications can be gleaned from the study including that the nonlinear decay transformation is effective at improving correlation perception in participants when the point encoding feature of size is used. While some limitations are present in the study, these are minimised by the study design and can be addressed in future research. Finally, future research can build upon the nonlinear decay transformation research to further improve correlation perception in people by assessing which point encoding feature is optimal. If this occurs, data visualisation research can be revolutionised and allow people to access enhanced scientific communication worldwide.

**ADD MORE LIT FOR THIS.**

**Appendix A**

There is much debate among researchers, including the creators of the emmeans package, concerning what is regarded as an appropriate sigma for effect size calculcations. The creators of emmeans do state that effect size results can widely differ due to different calculations in sigma and that their examples soberingly demonstrate the potential inaccuracy of effect sizes. The uncorrected effect sizes below were constructed using default coding within the eff\_size function detailed below (shown in r code).

eff\_size(emmeans\_contrasts, sigma = sigma(dot\_pitch model), = edf = df.residual(dot pitch model).

However, as one can see from the table below, the effect sizes are incredibly high even for nonsignificant findings. While significance cannot establish whether effects are present, it is important to judge the effects by select criteria and transparently report said criteria and rationale. Our rationale is three-fold. First, the effect sizes, in our opinion, are too high relative to other models that boast stronger claims for said effect sizes, e.g., the literacy model effect sizes. Second, regardless of it being possible, it is highly unlikely that these effect sizes are accurate; it would be incredibly rare. Third, some research suggests that this issue can be due to the calculation of sigma, the unreliability of a variable (dot pitch), and overestimated variance. This point means that the high variability or unreliability overestimates the SD and can cause inflated effect sizes. Thus, the correction below was used (shown in r code).

Tot\_sd = sqrt(random effect 1 sd + random effect 2 sd + residuals sd)

eff\_size(emmeans\_contrasts, sigma = tot\_sd, = edf = df.residual(dot pitch model).

This is a correction recommended by a few sources to counteract an evidently biased effect size. While this is not necessary for all effect size calculations for emmeans eff\_size, it can be used. With this correction, the resulting effect sizes, to our mind, were a) more consistent with other findings within the study, and b) more accurate. Ostensibly, these are effect sizes one would expect to find given the statistical results found. While the calculation is more conservative, it allows the dot pitch model results to be interpreted more stringently and consequently in relation to the other models. Finally, the unstandardised effect sizes (the differences in error rates) are shown and, some argue, are more useful than standardised effect sizes. This is because they retain more contextual information of the measured construct and are able to highlight factors such as deficiencies and sample specific information whereas standardised effect sizes lose this ability.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Contrast | Estimate | Z ratio | P value | Cohen’s d | Cohen’s d LCL | Cohen’s d UCL |
| Nonlinear – linear | 0.1128 | 3.184 | 0.0079 | 0.878 | 0.3377 | 1.419 |
| Nonlinear – inverted | 0.1894 | 5.325 | <.0001 | 1.475 | 0.9321 | 2.018 |
| Nonlinear – standard | 0.1690 | 4.775 | <.0001 | 1.316 | 0.776 | 1.8562 |
| Linear – inverted | 0.0776 | 2.152 | 0.1370 | 0.597 | 0.532 | 1.140 |
| Linear – standard | 0.0562 | 1.586 | 0.3867 | 0.438 | 0.103 | 0.9782 |
| Standard – inverted | 0.0205 | 0.0575 | 0.9396 | 0.159 | 0 | 0.3837 |