CLINICAL PSYCHOLOGY | RESEARCH ARTICLE

Reliable visual analysis of single-case data: A comparison of rating, ranking, and pairwise methods

Kevin R. Tarlow¹, Daniel F. Brossart²*, Alexandra M. McCammon², Alexander J. Giovanetti², M. Camille Belle² and Joshua Philip²

Abstract: The most common method of single-case data analysis is visual analysis, but interrater reliability among visual raters tends to be poor. A new paradigm of visual analysis is presented and tested with the goal of addressing this persistent limitation. In traditional visual analysis, graphs are viewed and rated one by one. However, in the ranking and pairwise comparison methods introduced here, graphs are compared to each other and sorted from least to most evidence of intervention effectiveness. Four visual raters scored a set of 30 previously published single-case graphs using a traditional rating method as well as the ranking and pairwise methods. As in previous studies of visual analysis, the raters failed to achieve acceptable interrater reliability with the traditional rating approach (α = 0.641). However, interrater reliability increased to satisfactory levels when graphs were scored with ranking (α = 0.847) and pairwise comparison (α = 0.860). Visual analysis scores based on the pairwise method were also used to evaluate the performance of three single-case effect size statistics.

Subjects: Psychological Methods & Statistics; Quantitative Methods; Research Methods in Education

ABOUT THE AUTHOR

This paper reflects Drs. Tarlow and Brossart’s ongoing interest in methodology issues related to single-case research. It is hoped that this paper benefits applied scientists doing single-case research by providing a way to efficiently conduct visual analysis with higher levels of reliability.

PUBLIC INTEREST STATEMENT

To evaluate treatment effects, clinicians and researchers often conduct a single-case experimental study with a baseline and treatment phase. Those results are then portrayed in a graph with the baseline and treatment phases plotted across time. This allows one to visually see the difference between those two phases and make decisions about the strength or nature of any treatment effect present. Unfortunately, this type of visual analysis typically results in poor reliability amongst visual raters. To address this limitation, three methods of visual analysis were examined. Interrater reliability increased to acceptable levels when raters sorted graphs using ranking and pairwise comparison methods. The pairwise method based on Elo’s method (1978) had the highest interrater reliability and requires less time than ranking. The results offer a promising direction for innovation in single-case visual analysis and may also aid in validating and interpreting single-case effect sizes when compared to visual analysis.

© 2021 The Author(s). This open access article is distributed under a Creative Commons Attribution (CC-BY) 4.0 license.
Keywords: single-case; single-subject; visual analysis; interrater reliability; effect size; Elo’s system

1. Introduction
Single-case experimental designs are used in many areas of psychological research, but investigators have yet to resolve substantial problems with single-case data analysis (Smith, 2012). The most common method of single-case data analysis is visual analysis; however, interrater reliability among visual raters tends to be poor. Manolov and Vannest (2019) noted that although visual analysis has historically been viewed as a gold standard for interpreting single-case data, it was often found to be unreliable. A meta-analysis by Ninci et al. (2015), the most comprehensive study of interrater reliability in single-case experimental designs, found visual raters were more reliable than earlier estimates (e.g. Danov & Symons, 2008; DeProspero & Cohen, 1979; Ottenbacher, 1993), but that on average raters failed to achieve “minimally acceptable agreement” (Ninci et al., 2015, p. 525). To be sure, visual analysis of single-case graphs is a demanding task in which raters must assess as many as six interdependent aspects of time-series data, including level, trend, overlap, variability, immediacy of effect, and consistency (Kratochwill et al., 2010). Regardless, single-case experimental methods cannot produce valid scientific knowledge if data analyses are unreliable.

Single-case investigators have proposed many strategies for improving interrater reliability in visual analysis. Visual aids, such as trend lines added to graphs, are among the most popular interventions for increasing interrater agreement (Lane & Gast, 2014; Manolov, 2017; Manolov & Vannest, 2019), but they do not yield statistically significant improvements in reliability (Ninci et al., 2015). Visual analysis training may improve reliability in some applications (e.g. Wolfe & Slocum, 2015a), though investigators have repeatedly demonstrated that expert raters familiar with single-case experimental methods have poorer interrater agreement than beginner raters (Ninci et al., 2015; Ottenbacher, 1993).

A host of statistical tools have been proposed as alternatives or supplements to visual analysis, though not without objection from visual analysis proponents (Baer, 1977; Carter, 2013). These single-case effect size statistics include nonparametric and nonoverlap methods (Parker, Vannest, Davis et al., 2011), parametric regression methods (Huitema & McKeen, 2000), multilevel modeling methods (Van den Noortgate & Onghena, 2007), Monte Carlo simulation methods (Borckardt et al., 2008; Tarlow & Brossart, 2018), and other methods. The large number of available effect size statistics has unfortunately created a new problem for single-case research: two investigators analyzing the same dataset with different statistical methods may arrive at different conclusions about the existence and size of intervention effects, similar to visual analysis (Smith, 2012). Single-case statistics can also be difficult to interpret without agreeing upon benchmarks for identifying small, medium, and large effect sizes (Parker et al., 2005).

In this study, we propose a new paradigm for visual analysis not previously explored, with the goal of developing a method that is sufficiently reliable for single-case research. Rather than constrain visual analysis to ratings of individual graphs, we tested ranking methods whereby sets of graphs were visually inspected and sorted from least to most evidence of experimental control or intervention effect. A ranking approach reframes the rater’s task from identifying the presence of an intervention effect to identifying the strongest evidence of intervention effect. A ranking approach, while distinct from a rating approach in its aims, nonetheless satisfies the needs of investigators who wish to identify the best evidence-based interventions, and who intend to identify “what [intervention], by whom, is most effective for this individual with that specific problem, and under which set of circumstances” (Paul, 1967). The following conceptual review of rating versus ranking systems supports our hypothesis that ranking tasks will yield better interrater reliability than rating tasks.
1.1. Rating versus ranking

Nearly all single-case visual analysis methods use a rating method in which graphs are inspected one at a time and assigned a score or descriptor based on the visual evidence of intervention effect. Effects are “small,” “medium,” “large,” or some other indicator of effect size. In many studies, raters judge only the presence or absence of experimental control (i.e. functional relationship between the intervention and outcome variable); the relative size of effect may not be of immediate interest. Rating is a subjective task, as evident by poor interrater reliability. One rater’s idea of “experimental control” or a “large” effect may differ in subtle but important ways from another rater.

With a ranking method, raters compare two or more single-case graphs to one another and arrange them in order from least to most evidence of intervention effect. Ranking tasks are potentially more reliable than rating tasks of a single graph, which rely on the rater’s unstandardized, idiosyncratic perceptions of experimental control and effect size. In a rating task, a graph is essentially evaluated against that rater’s conception of effect size or experimental control—raters may be thought to possess internal “mental graphs” that represent “small,” “medium,” and “large,” effects, and so on—and these internal metrics of comparison vary among raters. Put another way, while multiple raters may see the same target graph during a rating task, the “mental graphs” to which they compare the visualized graph differ from person to person. By comparison, ranking tasks minimize the need for (and hazard associated with) internal metrics of evaluation. In a ranking task, graphs are compared to each other—that is, the objects of comparison are external and standardized across raters. Although both methods are subjective, the ranking task provides a comparative context for evaluating the graphs. We therefore expect that a ranking procedure will yield better interrater reliability than a rating procedure.

1.2. Pairwise ranking with the Elo system

One limitation of ranking tasks is the difficulty of ranking large sets of items. Luckily, this challenge can be addressed by implementing a ranking system based on pairwise comparisons. According to Langville and Meyer (2012), “humans have a hard time ranking any set of items greater than size 5. Yet, on the other hand, we are particularly adept at pairwise comparisons” (p. 2). They pointed out that all scientific ranking systems are essentially based on pairwise comparisons. Sophisticated ranking systems are applied to a range of applied problems by aggregating many pairs of “this-or-that” selections—in areas as diverse as education, psychology, sports, and entertainment.

Elo (1978) introduced a popular ranking system based on pairwise comparisons. Elo’s system has since been adopted for many applications, and it forms the basis of the pairwise ranking system proposed in this study. The Elo system formulas are described in the Appendix, and its conceptual points are as follows:

1. Each graph is assumed to have an unknown “true” Elo score, which corresponds to the strength of the experimental intervention depicted
2. Two graphs are compared at a time; a visual rater selects a “winner,” which is the graph with the strongest evidence of intervention effect
3. After each matchup, the Elo scores of both graphs are recalculated; the winner’s score will increase and the loser’s score will decrease
4. Weaker graphs (i.e. graphs with lower pre-match Elo scores) are rewarded more for beating stronger graphs; stronger graphs are rewarded less for beating weaker graphs. The bottom line is that graphs depicting larger changes between phases, would be ranked higher than those with smaller differences between phases.
Elo’s rating system has a “simple elegance” that makes it a “near perfect way to rate and rank things by simple ‘this-or-that’ pairwise comparisons” (Langville & Meyer, 2012, p. 65). With each additional matchup, the Elo scores of graphs are expected to converge on their true scores, which in this case corresponds to the relative strength of their intervention effect as determined by a visual analyst.

1.3. Research questions and hypotheses

We predicted that ranking tasks would yield better interrater reliability for single-case visual analysis. We tested this hypothesis by comparing interrater reliabilities for three visual analysis tasks: rating, ranking, and pairwise comparison (using the Elo system). For the third task, given the iterative nature of the Elo system, we also planned to determine how many pairwise comparisons were necessary to achieve acceptable interrater reliability. In addition, we explored how well three single-case effect size statistics agree with the most reliable method of visual analysis, and we identified preliminary effect size benchmark values that corresponded with strong visual evidence of intervention effect.

2. Methods

2.1. Studies selected for inclusion

Smith (2012) conducted a systematic review of single-case studies published between 2000 and 2010, which included 409 studies from a range of fields. Thirty studies were randomly selected from the 409 articles included in Smith’s review. For selected studies that included more than one outcome variable or more than one participant, one time-series was randomly selected for inclusion. This yielded a sample of 30 previously published time-series datasets selected for visual analysis. As with the Smith review, the 30 graphs selected for this study included interventions designed to address a variety of developmental and psychological concerns, including autism, emotional and behavioral disorders, learning disabilities, intellectual disabilities, attention problems, anxiety, and trauma.
2.2. Graph standardization
Data points from the 30 selected time-series were digitally extracted with PlotDigitizer (2020). For designs with more than two phases (e.g. ABA, ABAB), only the first AB phase contrast was digitized in order to standardize the visual analysis task for all graphs. Time-series were then re-graphed with R (R Core Team, 2017) and printed onto 8.5-by-5.5 inch cards. To make them as visually uniform as possible, graphs were presented without visual aids (such as trend lines) and without y-axis scales or labels. An example of the standardized graph cards used in Tasks 1 and 2 is illustrated in Figure 1. The digitally extracted datasets and stimulus cards for all 30 graphs may be viewed and downloaded from a public data repository (data available at https://dataverse.tdl.org/dataverse/reliable_visual_analysis).

2.3. Visual raters
Four of the six authors served as raters and completed all rating and ranking tasks. Three of the four raters were psychology graduate students; one was an undergraduate student. Prior to this study, none of the raters were familiar with single-case experimental designs nor did they have any single-case research experience. All raters completed an online visual analysis tutorial (Wolfe & Slocum, 2015b) to standardize their approaches to the visual analysis tasks described below. The tutorial focused on correctly identifying level-change and slope-change patterns in single-case data. The tutorial’s creators found participation increased trainees’ percentage of correct visual ratings (Wolfe & Slocum, 2015a).

2.4. Visual analysis tasks
2.4.1. Task 1: rating
First, raters completed a simple rating of each standardized graph card. Intervention effects were scored on a scale of 0 to 100 in response to the prompt, “How certain are you that the intervention yielded an effect?” (similar to the procedure used in previous studies of visual analysis interrater reliability; DeProspero & Cohen, 1979; Kahng et al., 2010). A score of 0 designated “not at all
certain,” and 100 indicated “very certain”. Thirty graph cards were presented to raters one at a time, and the order of graphs was randomized for each rater by shuffling the cards.

2.4.2. Task 2: ranking
The visual raters then ranked all 30 graph cards from “least certainty that the intervention yielded an effect” to “most certainty that the intervention yielded an effect” (i.e. ranked 1 to 30). Raters were not given additional instructions or restrictions for how to approach the ranking task.

2.4.3. Task 3: pairwise comparisons
In the third task, raters were presented with random pairs of graphs and prompted to select the one with greater evidence of an intervention effect. All possible pairs of the 30 graphs were ranked in this way, with a total of 435 pairings. A computer program written in R randomized the pairs for each rater and displayed the graphs side-by-side on screen for comparison. The R program generated a record of each rater’s selections, which were input by clicking on the graph with greater evidence of intervention effect in each pair. A screenshot of the program used to record raters’ pairwise rankings is presented in Figure 2. The record of “this-or-that” selections was then used to calculate a set of 30 Elo scores for each rater (see Appendix for Elo system formulas). Elo scores were calculated using syntax written in R.

2.5. Measurement of interrater reliability
Interrater reliability is a broad concept, but can be conceptualized as the degree of agreement among independent judges when categorizing, ranking, or rating a set of objects (e.g. graphs). In this study, each judge produced one set of 30 ratings (Task 1), one set of 30 rankings (Task 2), and one set of Elo scores (Task 3). Krippendorff’s (2012) α was used to calculate the degree of interrater reliability on each of the three tasks. Krippendorff’s α is a general statistical measure of interrater agreement that may be applied to data of different levels (e.g. categorical, rank, interval). According to Krippendorff, acceptable interrater agreement requires that α > 0.80, though 0.67 < α < 0.80 may be tentatively interpreted. The interrater reliability package “irr” for R was used to calculate α (Gamer et al., 2012).

For Task 3, we also examined how many pairwise comparisons were necessary to achieve an acceptable level of interrater reliability (α > 0.80). Unlike some other pairwise rating systems, the Elo system does not require all possible pairwise comparisons to assign scores to each graph. With Elo’s method, each graph is assigned a starting score (zero) before the ranking procedure, and scores are iteratively adjusted following each pairwise comparison—in this case, Elo scores were re-calculated 435 times (each individual graph’s Elo score was re-calculated 29 times). We predicted that raters would achieve acceptable interrater reliability in Task 3, and that it would not require rating all possible pairwise comparisons to do so. We therefore calculated interrater reliability following each iteration of the 435 steps of the pairwise ranking task to determine the point at which raters achieved acceptable interrater reliability.1

2.6. Single-Case effect size statistics for comparison
Average Elo scores for each of the 30 graphs (calculated by averaging the four raters’ Elo scores) were compared via correlation to the estimated effect sizes obtained from three single-case statistics described below. These three effect sizes were selected to represent different methodological approaches to the statistical analysis of single-case data, ranging from simple to complex.

2.6.1. Percentage of Nonoverlapping Data (PND)
PND is a nonoverlap statistic, calculated as the percentage of intervention phase data points which exceed the maximum score in the baseline, i.e., the percentage of nonoverlap between phases (Scruggs et al., 1987). PND is among the most widely used single-case statistical metrics due to its
ease of use (Schlosser et al., 2008), though it has been criticized for failing to control baseline phase trend and for its sensitivity to baseline phase outliers (Allison & Gorman, 1993; Wolery et al., 2010). Several studies found PND to agree well with visual analysis (Ma, 2006; Scruggs & Mastropieri, 1994; Wolery et al., 2010). PND was calculated by hand for each of the 30 sampled datasets.

2.6.2. Baseline Corrected Tau
Baseline Corrected Tau (Tarlow, 2017a) is a nonparametric method based on Kendall’s (1962) rank correlation statistic and adapted from the popular Tau-U single-case statistic (Brossart et al., 2018; Parker, Vannest, Davis et al., 2011). The Baseline Corrected Tau effect size describes the association between intervention phase and outcome, controlling for baseline trend (if trend is present and statistically significant). Baseline Corrected Tau was developed to address several limitations of Tau-U, including Tau-U’s failure to constrain correlation values between −1 and +1 and its inadequate control of baseline trend. Unlike PND, Baseline Corrected Tau controls for the effect of baseline phase trend if baseline trend is statistically significant. An online calculator (Tarlow, 2017b) was used to calculate Baseline Corrected Tau for each of the 30 sampled datasets.

2.6.3. Interrupted Time-Series Simulation (ITSSIM)
ITSSIM modeling assumes that one observed time-series dataset could be explained by many plausible intervention effects and experimental conditions (Tarlow & Brossart, 2018). ITSSIM uses Monte Carlo simulation methods to determine which of those plausible models is most likely. Like Baseline Corrected Tau, ITSSIM controls for baseline trend, and it also models the effects of autocorrelation and heterogeneity between phases. ITSSIM yields several equivalent effect size statistics, including Cohen’s $d$, Pearson $r$, and $R^2$. For this study, ITSSIM $r$ values were calculated for each of the 30 sampled datasets using ITSSIM software (Tarlow, 2018). ITSSIM $r$ is interpreted as

![Figure 3. Interrater reliability (Krippendorff’s α) of Elo scores during the pairwise comparison task. At the end of the task, the interrater reliability for four raters was $\alpha = 0.86$, following all possible pairwise comparisons of 30 graphs ($n = 435$). However, interrater reliability was acceptable ($\alpha = 0.80$) after approximately 50% of the pairwise comparisons.](image-url)
the association between intervention and outcome, accounting for level-, trend-, and variability-change as well as autocorrelation.

3. Results

3.1. Comparing rating, ranking, and pairwise methods

The result of primary interest in this study was the relative interrater reliability coefficients among four judges on each of three visual analysis tasks. In Task 1, the raters examined 30 graphs individually and rated each on a scale of 0 to 100, with score indicating each judge’s certainty about a visibly detectable intervention effect. For Task 1, interrater reliability was $\alpha = 0.641$; this does not meet the criterion for acceptable agreement of $\alpha > 0.800$ (nor does it meet the threshold for tentative interpretation, $\alpha > 0.667$). In Task 2, the raters ranked 30 graphs from least to most certainty about a visibly detectable intervention effect. For Task 2, interrater reliability was $\alpha = 0.847$, indicating an acceptable level of agreement. In Task 3, Elo scores were calculated from raters’ pairwise comparisons of graphs; when presented with two graphs, raters selected the one with greater visible evidence of intervention effect. For Task 3, after raters completed all 435 possible pairwise comparisons, interrater reliability was $\alpha = 0.860$, the highest level of agreement across the three tasks. Raters took approximately 10–20 minutes to complete each of the three visual analysis tasks, and no task was substantially shorter or longer to complete on average.

In addition to calculating the final Elo score interrater following the pairwise comparisons of all 435 graphs, the interrater reliability was calculated following each of the 435 iterations of the pairwise comparison task. Figure 3 presents the iterative interrater reliabilities. As stated above, following 100% of all pairwise comparisons ($n = 435$), $\alpha = 0.860$. However, raters achieved acceptable agreement ($\alpha > 0.800$) after approximately half of the pairwise comparisons ($n = 230$). It is also noted that the order in which pairs of graphs were presented to each rater was randomized, so Elo scores were calculated in a different order across raters.

3.2. Visual versus statistical analysis

Three visual analysis scores were calculated for each graph by averaging the raters’ ratings, rankings, and pairwise (Elo) scores. The average rating, ranking, and pairwise scores were then correlated with each other and with the effect size estimates from three single-case statistics: PND, Baseline Corrected Tau, and ITSSIM r. PND effect sizes ranged from 0.00 to 1.00 (percentages were transformed to a zero-to-one scale); Baseline Corrected Tau effect sizes ranged from 0.07 to 0.88; ITSSIM r effect sizes ranged from 0.16 to 0.98.2 The correlation matrix for visual and statistical analyses is presented in Table 1. All correlations were statistically significant ($p < .001$) and they ranged from 0.63 to 0.99, indicating strong agreement between all visual and statistical analyses.

Correlations ranged from 0.90 to 0.99 among the average visual analysis scores, suggesting very strong agreement between rating, ranking, and pairwise methods on average, though only the ranking and pairwise methods had acceptable interrater reliability. The ranking and pairwise variables also had stronger correlations with the three single-case statistics than the rating variable. Comparing the three effect size statistics, Baseline Corrected Tau was most consistent with the three visual analysis tasks ($0.67 < r < 0.81$), PND was the second-most consistent with visual analysis ($0.64 < r < 0.76$), and ITSSIM r was relatively less consistent with visual analysis ($0.63 < r < 0.73$), though differences between statistical methods were small.

To identify preliminary interpretive benchmarks for effect size statistics based on their association with visual analysis, graphs were divided into four quartiles based on their average Elo scores (average Elo scores and graph data are available to view or download from a public data
Table 1. Correlations for visual and statistical analyses of 30 graphs

|       | (1) | (2) | (3) | (4) | (5) | (6) |
|-------|-----|-----|-----|-----|-----|-----|
| (1) Visual: Rating                                      |     |     |     |     |     |     |
| (2) Visual: Ranking                                     | .90 |     |     |     |     |     |
| (3) Visual: Pairwise (Elo)                              | .90 | .99 |     |     |     |     |
| (4) PND                                                 | .64 | .76 | .76 |     |     |     |
| (5) Baseline Corrected Tau                             | .67 | .80 | .81 | .93 |     |     |
| (6) ITSSIM Tau corrected r                              | .63 | .73 | .70 | .83 | .87 |     |

Note: All correlations are statistically significant, p < .001

Table 2. Visual analysis quartiles and corresponding effect size ranges

|                  | Q1     | Q2     | Q3     | Q4     |
|------------------|--------|--------|--------|--------|
| Elo score percentile | 0-25th | 25-50th | 50-75th | 75-100th |
| PND              | 0.00-0.74 | 0.00-0.80 | 0.82-1.00 | 0.86-1.00 |
| Baseline Corrected Tau | 0.07-0.61 | 0.15-0.68 | 0.52-0.81 | 0.71-0.88 |
| ITSSIM Tau corrected r     | 0.22-0.83 | 0.16-0.95 | 0.47-0.98 | 0.68-0.98 |

repository: https://dataverse.tdl.org/dataverse/reliable_visual_analysis. The range of effect size values within each of these four quartiles was then identified, with the aim of determining the effect sizes which correspond to strong visual evidence of intervention effect. Visual analysis scores and corresponding effect size values are presented in Table 2. For all three statistics, the first and second quartile ranges were largely overlapping, suggesting that these effect size statistics did a poor job of discriminating between graphs with relatively less visual evidence of intervention effect. ITSSIM r had particularly poor sensitivity, with values as high as r = 0.83 found in the bottom quartile of graphs. All graphs where PND > 0.80 and Baseline Corrected Tau >0.70 were in the top half of graphs when sorted by Elo score. All but one graph where ITSSIM r > 0.85 were in the top half of graphs when sorted by Elo score.

The scatterplots in Figure 4 illustrate the agreement between pairwise visual analysis and the ranking and statistical methods. It is immediately apparent from Figure 4 that the Elo pairwise visual analysis method is very consistent with visual analysis by ranking. Again, it is also apparent that ITSSIM r is least effective at discriminating between graphs with relatively strong and weak visual evidence of intervention effect.

4. Discussion

The purpose of this study was to evaluate a new paradigm of visual analysis for single-case experimental designs, with the goal of improving upon historically poor interrater reliability in single-case visual analysis. We predicted that traditional rating methods in which graphs are evaluated individually would yield poorer interrater reliability than ranking methods in which graphs are compared to one another and sorted. Consistent with this hypothesis, when raters were asked to visually rate 30 published single-case graphs individually, they failed to achieve an
acceptable level of interrater reliability, similar to previous studies of interrater reliability in visual analysis. However, interrater reliability increased to acceptable levels when raters were instead asked to sort the 30 graphs using ranking and pairwise comparison methods. The pairwise method, based on Elo’s (1978) rating system, had the highest interrater reliability of the three visual analysis methods. The pairwise method also had the advantage of requiring less cognitive effort from raters when compared to the full ranking method, as raters using the pairwise method had to consider only two graphs at a time in a series of “this-or-that” comparisons. Notably, high interrater reliability was achieved for the ranking and pairwise methods without the visual aids typically available to raters, such as y-axis labels, trend lines, or variability bands (Lane & Gast, 2014; Manolov, 2017; Manolov & Vannest, 2019); adding these aids could further improve interrater reliability. The graphs used in this study were also extracted from previously published single-case studies of real participants—in contrast to the less realistic computer-simulated graphs used in many studies of visual analysis reliability (e.g. DeProspero & Cohen, 1979; Kahng et al., 2010; Lieberman et al., 2010), which are typically more uniform and less ambiguous than graphs from real participant data.
These results, though preliminary, offer a new and promising direction for innovation in single-case visual analysis. Historically, despite visual analysis's status as the "gold standard" for single-case research, investigators could not be certain whether their analyses reflected the true effects of their interventions or merely the unreliable judgments of raters. This study's findings suggest there is an alternative approach to single-case visual analysis that yields more reliable results. It may be that visual raters are in fact quite reliable, contrary to conventional wisdom, as long as they are given the correct visual analysis task. When visual raters in this study were asked to rank a pair or set of single-case graphs against one another—rather than rate individual graphs—their rankings were highly consistent. If this result is successfully generalized beyond this study, single-case visual analysis may finally serve single-case investigators as a true gold standard for intervention assessment. Or, as Gwet (2014) stated, "If interrater reliability is high then raters can be used interchangeably without the researcher having to worry about the [rating] being affected by a significant rater factor" (p. 4).

This new approach to visual analysis also has important implications for the development and evaluation of single-case statistical methods. There are many ways to calculate effect size statistics with single-case data, and as a result different statistics applied to the same dataset may yield different conclusions about the existence and magnitude of an intervention effect. Investigators must therefore determine what statistics provide valid results, and under what circumstances. One useful way of validating single-case statistics is comparing their effect size coefficients to ratings from visual analysis (e.g. Brossart et al., 2006; Wolery et al., 2010). However, this approach to evaluation is only as useful as visual analysis is reliable. To say that an effect size statistic agrees with visual raters on average, when those same raters fail to agree with each other consistently, does not demonstrate the statistic's utility. Indeed, the proliferation of single-case effect size statistics is partly a result of visual analysis's poor interrater reliability (Campbell & Herzinger, 2010). Investigators with reliable methods of visual analysis will therefore be better equipped to identify the most useful statistical tools for single-case data analysis. Reliable visual analysis will also help investigators interpret statistical effect size values, as single-case statistics often yield coefficients that exceed typical interpretive benchmarks (Parker et al., 2005).

We demonstrated how reliable visual analysis scores (i.e. the Elo scores based on raters' pairwise comparisons of graphs) could be used to evaluate and interpret three single-case statistical methods (PND, Baseline Corrected Tau, and ITSSIM r). We first examined the correlations between the Elo scores and the three sets of effect sizes. All three effect size measures had statistically significant correlations with visual analysis, though Baseline Corrected Tau was most strongly correlated \(r = .81\), followed by PND \(r = .76\) then ITSSIM \(r = .71\). We then identified interpretive benchmarks for the statistics based on visual raters' Elo scores. The weakest intervention effects identified by visual raters (i.e. the 50% of graphs with smallest Elo scores) corresponded to the ranges PND < 0.80 and Baseline Corrected Tau < 0.70; ITSSIM r did not yield a clear benchmark.³ Interventions with PND > 0.80 or Baseline Corrected Tau > 0.70 could therefore reasonably be interpreted as “effective,” assuming the study meets the necessary design standards (e.g. Kratochwill et al., 2010) and statistical assumptions (e.g. Tarlow, 2018; Tarlow & Penland, 2016). Other interpretive benchmarks for these statistics may also be forthcoming with further study. Future visual analysis research utilizing ranking or pairwise methods could fruitfully extend this approach to evaluate other effect size statistics as well.

This study had several limitations. Only one pairwise rating method, the Elo system, was tested, though many other pairwise rating systems exist (Langville & Meyer, 2012); it is plausible that another rating system would produce even more reliable visual analysis scores. This study also only sampled 30 single-case graphs from Smith’s (2012) systematic review of published single-case research. These graphs included many types of interventions, and a sample of graphs limited to a specific type of intervention or participant population might yield different results, including different effect size benchmarks. That said, we feel the inclusion of many types of intervention graphs presented the most difficult challenge to visual
raters, and therefore makes raters’ consistency all the more encouraging. Even so, it should be noted that if one used a set of graphs that showed minimal to no treatment effects, the graphs could still be ranked, but the effect sizes would reflect the minimal treatment effects. This study also used only the first AB phase contrast in each graph, discarding additional phases from the studies with reversal designs (e.g. ABA, ABAB), in order to standardize the visual rating tasks. Including complex designs, or a mix of simple and complex designs, may affect the reliability of the rating task. In addition, all raters completed the three rating tasks in the same order (rating, ranking, pairwise comparison), so the influence of order effects cannot be ruled out; however, there was an interval of at least one week between each task in order to minimize this threat to validity.

In conclusion, it is recommended that investigators using single-case experimental designs consider ranking and pairwise comparison methods of visual analysis when designing their studies, instead of (or in addition to) traditional rating methods in which graphs are examined individually. The results of this study suggest that replacing traditional rating procedures with visual analysis tasks in which graphs are sorted or ranked may greatly improve the reliability between visual raters. Pairwise methods in particular may be an attractive option for raters because they may be applied to any number of graphs, whereas ranking approaches become cumbersome as the number of graphs increases. Furthermore, our results suggest raters may not need to score all possible pairs of graphs; scoring as few as half of all possible pairs may be sufficient to achieve good interrater reliability. Investigators who wish to utilize a rank-based visual analysis method, but otherwise lack the ability to replicate the procedure outlined in this article, can access all graphs and Elo score data for this study from a public data repository (https://dataverse.tdl.org/dataverse/reliable_visual_analysis). Investigators may estimate an approximate Elo score (and visual analysis percentile score) for their graph by visually comparing it with the graphs used and scored in this study. Single-case statistical methodologists could also use this public data to evaluate effect size statistics other than the three coefficients included in this study.

Funding
The authors received no direct funding for this research.

Author details
Kevin R. Tarlow1
E-mail: krtarlow@gmail.com
Daniel F. Brossart2
E-mail: brossart@tamu.edu
E-mail: abulovas@tamu.edu
ORCID ID: http://orcid.org/0000-0003-2619-1450
Alexandra M. McCammon2
E-mail: abulovas@tamu.edu
E-mail: giovanetti@tamu.edu
M. Camille Belle2
E-mail: cbelle@tamu.edu
Joshua Philip1
E-mail: joshphilip14@tamu.edu
1 Department of Psychology, Hampton University, Hampton, United States.
2 Department of Educational Psychology, Texas A&M University, College Station, United States.

Citation information
Cite this article as: Reliable visual analysis of single-case data: A comparison of rating, ranking, and pairwise methods, Kevin R. Tarlow, Daniel F. Brossart, Alexandra M. McCammon, Alexander J. Giovanetti, M. Camille Belle & Joshua Philip, Cogent Psychology (2021), 8: 1911076.

Notes
1. In order to examine the increasing interrater reliability during the iterative pairwise comparison task, it was necessary to set each graph’s starting Elo score value at a random normal deviate very close to zero, with μ = 0 and s = .001. This adjustment had no effect on final Elo scores, but permitted the starting interrater reliability to be α = 0. If all Elo scores were initially set at exactly zero, the starting interrater reliability would indicate perfect agreement, α = 1.

2. The absolute values of negative effect size coefficients were used in all analyses. The direction of effect size coefficients (positive or negative) typically reflected the measurement design rather than the intervention effectiveness. For example, interventions designed to decrease an unwanted outcome would yield a negative effect size coefficient whereas interventions designed to increase a desired outcome would yield a positive effect size coefficient. In this study, only the relative magnitudes of effect sizes were of interest, and using absolute values of all effect size coefficients aided interpretation.

3. The range ITSSIM r < 0.85 could tentatively be interpreted the same way, though one graph from the bottom 50% of Elo scores had the effect size ITSSIM r = 0.95.

Data availability statement
The data that support the findings of this study are openly available in [repository name e.g. figshare] at https://dataverse.tdl.org/dataverse/reliable_visual_analysis.

References
Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. Behaviour Research and Therapy, 31(6), 621–631. https://doi.org/10.1016/0005-7967(93)90115-8
Baer, D. M. (1977). Perhaps it would be better not to know everything. Journal of Applied Behavior Analysis, 10(1), 167–172. https://doi.org/10.1901/jaba.1977.10-167

Borkardt, J. J., Nash, M. R., Murphy, M. D., Moore, M., Shaw, D., & O’Neil, P. (2008). Clinical practice as natural laboratory for psychotherapy research: A guide to case-based time-series analysis. American Psychologist, 63(2), 77–95. https://doi.org/10.1037/0003-066X.63.2.77

Brossart, D. F., Laird, V. C., & Armstrong, T. W. (2018). Interpreting Kendall’s Tau and Tau-U for single-case experimental designs. Cogent Psychology, 5(1), 1518687. https://doi.org/10.1080/23311908.2018.1518687

Brossart, D. F., Parker, R. I., Olson, E. A., & Mahadevan, L. (2006). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. Behavior Modification, 30(5), 531–563. https://doi.org/10.1177/01454455060145445508318606

Campbell, J. M., & Herzinger, C. V. (2010). Statistics and single subject research methodology. In D. L. Gast (Ed.), Single subject research methodology in behavioral sciences (pp. 417–453). Routledge.

Carter, M. (2013). Reconsidering overlap-based measures for quantitative synthesis of single-subject data: What they tell us and what they don’t. Behavior Modification, 37(3), 378–390. https://doi.org/10.1177/0145445513476509

Donov, S. E., & Symons, F. J. (2008). A survey evaluation of the reliability of visual inspection and functional analysis graphs. Behavior Modification, 32(6), 828–839. https://doi.org/10.1177/0145445508318606

DeProspero, A., & Cohen, S. (1979). Inconsistent visual analyses of intrasubject data. Journal of Applied Behavior Analysis, 12(4), 573–579. https://doi.org/10.1901/jaba.1979.12-573

Elo, A. E. (1978). The rating of chess players, past and present. Arco Publishing, Inc.

Gamer, M., Lemon, J., Fellows, J., & Singh, P. (2012). Irr: Various coefficients of interrater reliability and agreement. R package version 0.84. Institute for Statistics and Mathematics at Wirtschaftsuniversität Wien. Retrieved from https://CRAN.R-project.org/package=irr

Gwet, K. L. (2014). Handbook of interrater reliability (4th ed.). Advanced Analytics.

Huitema, B. E., & McKeen, J. W. (2000). Design specification issues in time-series intervention models. Educational and Psychological Measurement, 60(1), 38–58. https://doi.org/10.1177/001316440021970358

Kahng, S., Chung, K. M., Gutshall, K., Pitts, S. C., Kao, J., & Grolombok, K. (2010). Consistent visual analyses of intrasubject data. Journal of Applied Behavior Analysis, 43(1), 35–45. https://doi.org/10.1901/jaba.2010.43-35

Kendall, M. G. (1962). Rank correlation methods (3rd ed.). Hafner.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shodish, W. R. (2010). Single-case design technical documentation. Institute of Education Sciences. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/wwc_scd.pdf

Krippendorff, K. (2012). Krippendorff’s alpha. In N. J. Salkind (Ed.), Encyclopedia of research design (pp. 669–673). Sage.

Lane, J. D., & Gast, D. L. (2014). Visual analysis in single case experimental design studies: Brief review and guidelines. Neuropsychological Rehabilitation, 24(3–4), 445–463. https://doi.org/10.1080/0909062011.2013.815636

Longville, A. N., & Meyer, C. D. (2012). Who’s #1? The science of rating and ranking. Princeton University Press.

Lieberman, R. G., Yoder, P. J., Reichow, B., & Wolery, M. (2010). Visual analysis of multiple baseline across participant graphs when change is delayed. School Psychology Quarterly, 25(1), 28–44. https://doi.org/10.1037/a0018600

Mo, H. H. (2006). An alternative method for quantitative synthesis of single-subject researches: Percentage of data points exceeding the median. Behavior Modification, 30(3), 598–617. https://doi.org/10.1177/014544550427974

Manolov, R. (2017). Linear trend in single-case visual and quantitative analysis. Behavior Modification, 42(5), 684–706. https://doi.org/10.1177/0145445517726301

Manolov, R., & Vannest, K. J. (2019). A visual aid and objective rule encompassing the data features of visual analysis. Behavior Modification, 34(4), 510–541. https://doi.org/10.1177/0145445518854323

Ninc, J., Vannest, K. J., Willson, V., & Zhang, N. (2015). Interrater agreement between visual analysts of single-case data. Behavior Modification, 39(4), 510–541. https://doi.org/10.1177/0145445515581327

Ottenbacher, K. J. (1993). Interrater agreement of visual analysis in single-subject decisions: Quantitative review and analysis. American Journal on Mental Retardation, 98(1), 135–142.

Parker, R. I., Brossart, D. F., Vannest, K. J., Long, J. R., Garcia De-Alba, R., Baugh, F. G., & Sullivan, J. R. (2005). Effect sizes in single case research: How large is large? School Psychology Review, 34(1), 116–132. https://doi.org/10.1080/2796015.2005.12086279

Parker, R. I., Vannest, K. J., & Davis, J. L. (2011). Effect size in single-case research: A review of nine nonoverlap techniques. Behavior Modification, 35(4), 303–322. https://doi.org/10.1177/0145445511399147

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011). Combining nonoverlap and trend for single-case research: Tau-U. Behavior Therapy, 42(2), 284–299. https://doi.org/10.1016/j.beth.2010.08.006

Paul, G. L. (1967). Strategy of outcome research in psychotherapy. Journal of Consulting Psychology, 31(2), 109–118. https://doi.org/10.1037/h0024436

PlotDigitizer (2020). PlotDigitizer (v2.11) [Computer software]. https://plotdigitizer.com

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Retrieved from www.R-project.org/

Schlosser, R. W., Lee, D. L., & Wendt, O. (2008). Application of the percentage of nonoverlapping data (PND) in systematic reviews and meta-analyses: A systematic review of reporting characteristics. Evidence-Based Communication Assessment and Intervention, 2(3), 163–187. https://doi.org/10.1080/17489530802505412

Scruggs, T. E., & Mastropieri, M. A. (1994). The effectiveness of generalization training: A quantitative synthesis of single-subject research. In T. E. Scruggs & M. A. Mastropieri (Eds.), Advances in learning and behavioral disabilities (Vol. 8, pp. 259–280). Emerald.
Appendix

Elo System Formulas

The revised Elo rating method outlined by Langville and Meyer (2012) was used in this study. First, all 30 graphs were assigned an initial rating of \( r = 0 \). Two graphs, \( i \) and \( j \), were then randomly chosen and presented to the visual rater for comparison. Following the rater's selection of the graph with the greater evidence of intervention effect (i.e. the “winner”), new ratings were calculated as

\[
\begin{align*}
    r_{i(new)} &= r_{i(old)} + K(S_{ij} - \mu_{ij}) \\
    r_{j(new)} &= r_{j(old)} + K(S_{ij} - \mu_{ij})
\end{align*}
\]

The constant, \( K \), called the “K-factor,” determines how much rating scores adjust after each pairwise comparison, with higher \( K \) values leading to more volatile ratings. Here the K-factor was set to \( K = 32 \), following Langville and Meyer’s example. \( S_{ij} \) equals 1 if graph \( i \) is selected over \( j \), or it equals 0 if graph \( j \) is selected over graph \( i \). The final term, \( \mu_{ij} \), is determined by the logistic function

\[
\mu_{ij} = L(d_{ij}/\xi) = \frac{1}{1 + 10^{-d_{ij}/\xi}}
\]

where

\[
d_{ij} = r_{i(old)} - r_{j(old)}
\]

and \( \xi \) is a logistic parameter determining the spread of the ratings; here, \( \xi = 1000 \). This value indicates that a graph with a pre-matchup rating 1000 points higher than its opponent would be ten times more likely to be selected as the “winner” than the lower-rated graph. Each graph was paired with the other 29 graphs in random order, yielding a total of 435 pairwise comparisons.
