Original Article

When the Truth Hits You Between the Eyes

A Software Tool for the Visual Analysis of Single-Case Experimental Data

Isis Bulté and Patrick Onghena

Faculty of Psychology and Educational Sciences, Katholieke Universiteit Leuven, Belgium

Abstract. Visual data analysis is an important first step when evaluating intervention effects. This also holds for analyzing data from single-case experiments. Because most software packages do not offer customized facilities for constructing single-case graphs and are not particularly suited to perform single-case visual data analyses, we created an R package to help researchers in making graphical representations of single-case data and to transform graphical displays back to raw data. In addition to a basic plotting function, we included some tools to facilitate the use of three interpretative principles for visually analyzing single-case data: plotting a measure of *central location* as a horizontal reference line; displaying *variability* with (trimmed) range bars, range lines, and trended ranges; and displaying *trends* with a vertical line graph, by fitting a robust linear trend, or by plotting running medians. Finally, we included a function to extract raw data values from published graphs.

Keywords: single-case experiments, visual data analysis, software, R package

Visual analysis is the primary method of evaluating data from single-case experiments (Busk & Marascuilo, 1992; Kahng et al., 2010; Kazdin, 2011; Parsonson & Baer, 1992). Although several statistical alternatives have been suggested over the years (for an overview, see, e.g., Campbell & Herzinger, 2010; Gorman & Allison, 1997; Houle, 2009), these alternatives are only rarely used (e.g., Brossart, Parker, Olson, & Lakshmi, 2006; Carter, 2009; Kahng et al., 2010). When looking at trends in single-case research, Kratochwill and Brody (1978), Busk and Marascuilo (1992), and Long and Hollin (1995) indicated that in less than 10% of their sampled studies statistical analyses were used. Parker et al. (2005) reanalyzed published single-case data and found over 65% of the studies in their sample relying on visual analysis alone. Effect sizes, confidence intervals, and tests of statistical significance were found in only 11% of the articles. A reason for this difference with group studies, in which statistical analyses are most common, lies in the two conflicting traditions from which single-case and group studies arose (Allison, Franklin, & Heshka, 1992). Scientific group study research relies on making inferences based on formal statistical procedures, largely based on the work of Ronald Fisher and Jerzy Neyman. Single-case designs and applied behavior analysis have their roots in the work of Murray Sidman and Burrhus Frederic Skinner, who rejected statistical analysis and relied solely on visual analysis of the graphed data (Allison et al., 1992). This dominance of visual analysis in published single-case research does not necessarily mean that it is generally accepted to be the best data-analytic technique available. In past research, questions have been raised concerning the consistency (or reliability), the sensitivity, and the specificity of visual analysis.

Consistency

Visual analysis is often criticized for having no empirically established formal decision guidelines, leaving room for subjectivity and inconsistency. This is demonstrated by the average poor interrater agreement that is found in several studies (e.g., Bobrovitz & Ottenbacher, 1998; DeProspero & Cohen, 1979; Fisch, 1998; Jones, Weinrott, & Vaught, 1978; Ottenbacher, 1990; Park, Marascuilo, & Gaylord-Ross, 1990; Ximenes, Manolov, Solanas, & Quera, 2009). This discouraging low consistency was not confirmed by all researchers, with, for example, Kahng et al. (2010) reporting a high level of agreement among different judges when replicating the 1979 DeProspero and Cohen study. Potential reasons for these contradicting results are procedural differences in the instructions and response measures used, and the fact that the earlier results could be outdated because of more and better training opportunities that exist for judges nowadays. Also, the alarming results found in earlier studies could be partly explained by the difference between real-life settings and the artificial settings used in the studies (Brossart et al., 2006; Parsonson & Baer, 1992).

Several variables were shown to increase interrater agreement: Training judges in using a standard trend estimation procedure, prior knowledge of the participant of whom

DOI: 10.1027/1614-2241/a000042

the data are graphed, and the use of judgmental aids (Skiba, Deno, Marston, & Casey, 1989). The area of expertise seems to have an influence, with statistically trained judges showing more agreement than single-case analysts (Harbst, Ottenbacher, & Harris, 1991). Some characteristics of the data also play a role: changes in mean shift and level across phases seem to yield higher agreement among judges than changes in variability and slope (DeProspero & Cohen, 1979; Gibson & Ottenbacher, 1988; Knapp, 1983), whereas data overlap and variability between phases seems to have little influence (Gibson & Ottenbacher, 1988). The effect of serial dependency is not clear. It is thought that autocorrelation negatively influences agreement among judges (Jones et al., 1978; Matyas & Greenwood, 1990), but some studies only found weak effects (Gibson & Ottenbacher, 1988).

Sensitivity and Specificity

Proponents of visual analysis claim this method to be more conservative and less sensitive than statistical analysis (Kazdin, 2011). This insensitivity is seen as an advantage, because small treatment effects will be ignored and only very large effects, which are most likely also clinically (or practically) relevant, will be detected (e.g., Carter, 2009; Jones et al., 1978; Ottenbacher, 1990; Park et al., 1990; Parsonson & Baer, 1992). Overlooking consistent but relatively weak effects could however be problematic, for instance in developing new theories or technologies (Kazdin, 2011). Not all studies confirmed this claimed conservatism. Bobrovitz and Ottenbacher (1998) found a high agreement between the results of visual and statistical analyses, concluding that both methods are equally sensitive and specific. Fisch (1998) found very high Type I error rates, and Matyas and Greenwood (1990) and Normand and Bailey (2006) found that participants were more likely to detect nonexistent effects (Type I errors) than they were to omit existing effects (Type II errors).

Since findings up till now are inconsistent and contradictory, concerns about the error rates in visual analysis remain, and questions about the quality of the validity studies have been raised. One issue involves the unrealistic conditions that are used, because validity studies differ in the extent to which they resemble natural conditions. In real-life situations, visual analysts have to make their judgment in a particular context and can judge the degree of effectiveness of a given intervention instead of being forced to make a dichotomous yes/no decision. Besides the available response options, also characteristics of the judges (e.g., trained/untrained) and the way of graphically presenting the data (e.g., addition of trend lines) could be confounding factors and more research is definitely needed (Brossart et al., 2006; Carter, 2009; Furlong & Wampold, 1982; Parsonson & Baer, 1992; Rojahn & Schulze, 1985; Wampold & Furlong, 1981). Allison et al. (1992) state that Type I error rates could even be larger in real-life applications because of the use of response-guided experimentation. If each data point is judged instantly and separately to determine how to continue the experiment, these multiple analyses increase the probability of making at least one Type I error. Allison et al. (1992) therefore correct the estimate of 10% for Type I error rates in visual analysis to 25%. Ferron and Jones (2006) presented a method to control these Type I error rates, by using random assignment and a ‘blind’ data analyst.

A second issue concerns the validation criteria that are used in the studies. Some studies (e.g., Normand & Bailey, 2006) constructed data sets and graphs with known characteristics. However, many studies compare the results of visual analysis with those of statistical analysis, like the split-middle method (e.g., Ottenbacher, 1990; Richards, Taylor, & Ramasamy, 1997), time-series analysis (e.g., Jones et al., 1978), and randomization tests (e.g., Park et al., 1990) and different statistics tend to yield different conclusions. Since no consensus exists on which statistical test to use, without a power analysis none of them can be taken as the standard to evaluate the performance of visual judges (Brossart et al., 2006; Parsonson & Baer, 1992; Ximenes et al., 2009). For experiments with relatively large sample sizes, well-established procedures exist to determine the effects of the intervention. Provided that some assumptions about the population are met, parametric statistical tests (e.g., *t*-tests and ANOVAs) can be used validly. For small-*n* experiments the same procedures as for large-*n* designs do not necessarily apply because there are more doubts about the plausibility of the assumptions. The most important assumption that might be violated is that of serial independence (Kazdin, 2011). Autocorrelation might however also complicate the interpretations when visually analyzing single-case data, since it can be confounded with the existence of a treatment effect (e.g., Matyas & Greenwood, 1990; Rojahn & Schulze, 1985). Several statistical techniques were proposed for analyzing single-case data (for an overview, see, e.g., Campbell & Herzinger, 2010; Gorman & Allison, 1997; Houle, 2009), but all were criticized for various reasons (e.g., they pose limitations on the researchers, they tend to ignore clinical significance, . . .), and there is still some debate on which of them is suitable. Several effect size indices based on visual analysis criteria have also been developed (e.g., Ma, 2006; Parker, Hagan-Burke, & Vannest, 2007), but here too no clear guidelines exist yet on which measure to use and how to interpret it.

The Intraocular Trauma Test: Does the Result Hit You Between the Eyes?

Visual analysis is a necessary step when evaluating intervention effects. Even statisticians increasingly emphasize the importance of graphical data analysis. Wilkinson and the APA task force on Statistical Inference (1999) and Wilkinson (2005), for example, advise to first look at the data before computing any statistics. Most software packages, however, are not specialized in constructing single-case graphs. We hereby present a tool to aid researchers in making graphical representations of their single-case data, and transforming graphical displays back to raw data. To create the graphical functionalities, we used the programming environment R. This was chosen because, as an open source implementation of the S-PLUS language, it can be downloaded at no cost from the CRAN website (cran.r-project.org). R is a very powerful and flexible tool, which has very good graphical possibilities to obtain simple as well as complex visual displays of the data, and which is able to deal with unusual data sets and problems (Kelley, 2007). The R scripts that we created are available on our website (ppw.kuleuven.be/english/research/mesrg).

Visual Analysis: A Software Tool

Kazdin (2011) defines visual analysis as 'reaching a judgment about the reliability or consistency of intervention effects by visually examining the graphed data' (p. 286). Visual analysis depends on many characteristics of the data, but most importantly on the magnitude of changes across phases (differences in the mean level of performance and level shifts at the moment of phase change) and on the rate of those changes (changes in trend and the latency of changes). Generally speaking, three variables are taken into account: central location (and level), variability, and trend (Franklin, Gorman, Beasley, & Allison, 1997; Morley & Adams, 1991). Therefore we did not only create a basic plotting function to display data from single-case experiments, but we also incorporated some tools to facilitate the use of these three interpretative principles of visual analysis.

From Data to Graph

Single-case data are usually displayed by plotting the measure of time on the abscissa and the dependent variable on the ordinate, as displayed in Figure 1. For *single-case alternation designs* (left panel), in which the basic strategy consists of the rapid alternation of two or more conditions within a single case, the data points of each condition are connected, resulting in multiple lines in the same area of the plot. In *single-case phase designs* (middle panel), comparisons are made within a time series and the case's performance is evaluated over time across baseline (A) and intervention (B) phases. The graphical display connects the data points in each phase, yielding an interrupted line with a space between the phases. In this space, a vertical line is placed to mark the phase change. Some contrasting findings were published on the effects of these graphing conventions. Knapp (1983) found that judges were more conservative when the phases are not separated, whereas Carter (2009) found no substantive evidence that judgments were affected. If any effect, he found more Type I errors when the graphs are presented without phase change lines and when the data points across phase changes are connected. Because of the lack of a clear opinion and because these conventions are widely used, we decided to incorporate them in the graphical display of phase designs. Users who wish to change these settings can easily do so by slightly changing the R code. Data from *multiple baseline AB designs* (right panel), in which several AB phase designs are implemented simultaneously to different persons, behaviors, or settings, are displayed by plotting these different AB designs beneath each other, so that the staggered administration of the intervention becomes clear. As can be seen in Figure 1, standard labels are given to the conditions and phases ("A" and "B") and to the abscissa ("Measurement Times") and ordinate ("Scores"). These can be personalized within the code.

These basic single-case graphs can be obtained by calling the function `graph` ("`design`"), where with the design argument the user can specify which design was used in the experiment. Possibilities are a phase design ("`AB`," "`ABA`," "`ABAB`"), an alternating treatments design ("`ATD`"), and a multiple baseline AB design ("`MBD`"). A more detailed explanation of these different design types can be found in Barlow, Nock, and Hersen (2009). All the functions explained below will work for the above-mentioned design types. For illustration purposes we will focus on AB phase designs in what follows. The use of the general interpretative principles of central location, variability, and

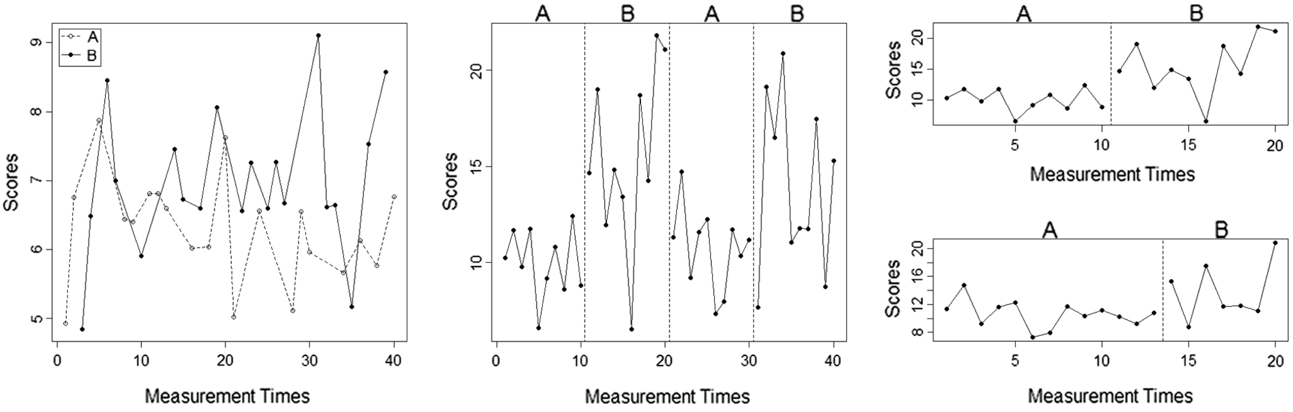


Figure 1. Graphical display of three hypothetical examples of single-case research design types. In the left panel an alternating treatments design with two conditions is shown (`graph(design = "ATD")`). The middle panel displays an ABAB phase design (`graph(design = "ABAB")`). And in the right panel an example of a multiple baseline design with two units is given (`graph(design = "MBD")`).

trend in the next paragraphs is largely based on Morley and Adams (1991) and Franklin et al. (1997).

Central Location

Changes in level across phases appear to be the characteristic that judges use most often to come to a decision and it is the characteristic that is associated with the highest degree of interrater agreement, rater certainty, and rater confidence (Bailey, 1984; Fisch, 1998; Furlong & Wampold, 1982; Gibson & Ottenbacher, 1988; Normand & Bailey, 2006; Parsonson & Baer, 1992; Wampold & Furlong, 1981). Central location can be incorporated in the single-case plot by superimposing a horizontal reference line on the raw time series. This will make treatment effects (demonstrated by differences in level) more visible, while also providing a basis for analyzing variability and trend. With the function `graph.CL(design,CL,tr)` the user can choose which measure of central tendency has to be plotted as a line parallel to the abscissa. An example of an AB phase design where the mean is plotted as a reference line for each phase is given in Figure 2. Note that there is no legend indicating which measure of central tendency is displayed. Users can mention this in the figure's caption. This recommendation also applies to most of the subsequent graphs.

The mean (`CL = "mean"`) is not resistant to the influence of outliers, so sometimes it could be better to consider the median (`CL = "median"`). However, in small samples the median may also not be very representative, because it only takes into account the one or two central data points of the whole data set. A possible solution here is to use the broadened median (`CL = "bmed"`), which is calculated based on the three, four, five, or six middle values of the data set, depending on the total number of data points (Morley & Adams, 1991). This way this measure is sensitive to a larger proportion of the data than the median, while also being robust to the influence of outliers.

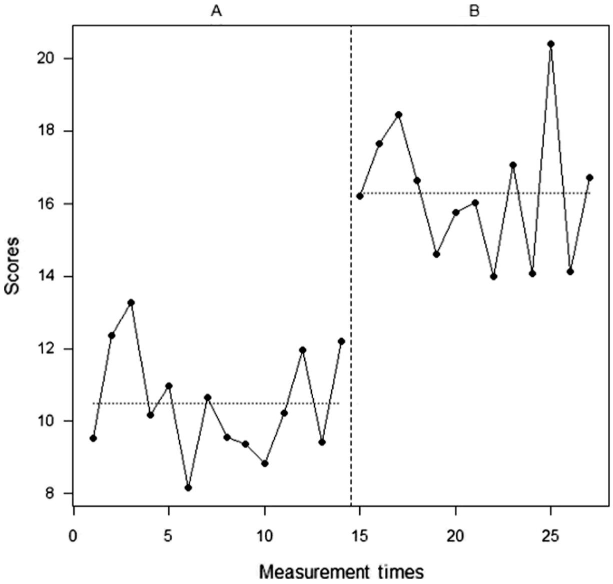


Figure 2. Hypothetical example of a possible display of central location for an AB phase design. The mean is plotted in each phase as a horizontal reference line superimposed on the raw time-series data (`graph.CL(design = "AB", CL = "mean")`).

The median, however, has a larger standard error than the mean when the population is normally distributed (Wilcox, 2005). It is also strongly affected by the observations in the center of the distribution, which is not the case for the mean. Another way of dealing with the lack of robustness of the mean as a measure of central tendency is by discarding the observations in the tails of the distribution (the extreme values) and calculating the so-called "trimmed mean" on the remaining observations (`CL = "trimmean"`). The percentage of observations that has to be removed from each end of the distribution before computing the mean can be set with the "`tr`" argument, and can be any value from 0 (regular arithmetic mean) to 0.5. Usually 20% of the observations is trimmed (so `tr = 0.2`).

While the trimmed mean removes a fixed proportion of the observations, whether they are outliers or not, Huber's M-estimators of location (Huber & Ronchetti, 2009) first evaluate each observation to determine if it actually is an outlier compared to the rest of the data and then give less weight to those outlying values. For this evaluation a constant K needs to be specified that can have any value between 0 and ∞, and which actually comes down to balancing between robustness and efficiency: when K = 0, the M-estimator equals the median (maximally robust), while when K→∞ it equals the arithmetic mean (maximally efficient) (Jeng, 2010). Therefore, M-estimators of location could be considered as a generic measure of central tendency of which the mean and the median are special cases. Usually a percentile of the standard normal distribution is chosen as the constant K. Wilcox (2005) suggests using K = 1.28, which corresponds to the 90th percentile of the standard normal distribution and covers 80% of the underlying distribution. When choosing a larger value for K, the coverage gets larger until the whole distribution is covered when K→∞ (as with the mean). In other words, the larger K gets, the wider the range of plausible observations becomes, and the lesser observations will be classified as outliers. To display this generic measure of central tendency, the CL argument should be put to "`mest`," and the desired value for the constant should be specified with the `tr` argument (e.g., `tr = 1.28`). The function mest(x,bend = 1.28) from Wilcox (2005) is used for the calculation of this measure. In his book, Wilcox (2005) provides R functions for several other robust estimators of location. These estimators could also be included by making slight adjustments to the R code.

Variability

Relative variation within and between phases is often overlooked or ignored by visual analysts. This seems to be influenced by the area of expertise of the judges: Whereas statistically trained judges take into account the type of the effect and the amount of variability across phases, experts in visual analysis tend to look only at the magnitude of the effect (Furlong & Wampold, 1982; Wampold & Furlong, 1981).

However, variability in the data could have an influence on the decisions made by the judges, with a higher degree of variability being related to higher Type I error rates (Matyas & Greenwood, 1990). By graphically displaying some estimates of variability, we want to separate variability from other aspects in the display, like level and trend with which it is often confounded. This way, we aim at making it easier to attend to variability when visually analyzing the data.

The function `graph.VAR(design,VAR,dataset,CL,tr)` was developed for this purpose. It includes two commonly used methods to demonstrate variability: range bars and range lines. *Range bar graphs* (`VAR` = "`RB`") consist of a vertical line for each phase, created by connecting three points: An estimate of central tendency, the minimum, and the maximum, while the raw data points are usually not plotted. Depending on the desired measure of central tendency, the `CL` argument should be put at "`mean`," "`median`," "`bmed`," "`trimmean`," or "`mest`" (with the additional `tr` argument for the trimmed mean and the M-estimator). In the upper left panel of Figure 3, a range bar graph with the mean as a measure of central tendency is displayed.

Range lines (`VAR` = "`RL`") are superimposed on the raw data and therefore do include information about the timing of the data points. This way information about variability is displayed, while at the same time highlighting possible trends. A pair of lines is drawn parallel to the *X*-axis, and passing through the lowest and the highest values for each phase. An example is given in the upper middle panel of Figure 3.

One problem with these two methods of displaying variation is that the range is severely influenced by the presence of outliers. This may be overcome by using a *trimmed range* (`dataset` = "`trimmed`"), in which only a sample of the data is used. The original data set is trimmed with 10–20% by taking out the most extreme values (i.e., the lowest and highest data points). This is demonstrated for range bars (lower left panel) and range lines (lower middle panel) in Figure 3. Notice that for range bars, the omitted values are displayed as dots above and below the vertical line. Two other problems with range bars and range lines are that variability may be confounded with trend and that changes in variability within phases are not displayed. This could be solved by plotting a trended range (`VAR` = "`TR`"), as demonstrated in the upper right panel of Figure 3. For each phase two lines are drawn, displaying possible changes in variability. First the phase is divided into two halves along the *X*-axis. Then the middle time point is determined for each phase half, and the minimum and maximum values of the dependent variable within that half are plotted at that time point. Finally the minimum values of each half are connected, as well as the maximum values, which results in two lines. Trended ranges can also be applied to the data after the lowest and the highest values in each half of the data set have been trimmed, by setting the `dataset` argument to "`trimmed`" (lower right panel).

Trend

Whereas level changes between phases are the characteristic associated with the highest degree of interrater agreement, changes in trend are the characteristic that is most often associated with inconsistent interpretations (Bailey, 1984; Gibson & Ottenbacher, 1988; Parsonson & Baer, 1992). Judges seem to focus primarily on shifts in level, being unable to differentiate these from trends in the data, with similar difficulties in detecting trends for experienced and non-experienced judges (Fisch, 1998; Furlong & Wampold, 1982; Wampold & Furlong, 1981). After training judges in using a standard trend estimation procedure, interrater agreement as well as confidence in the judgments increased. However, training does not necessarily lead to better performance (Knapp, 1983; Richards et al., 1997). It can result in more conservative judgments or make analysts rely only on those criteria in which they have been trained. Besides training, the use of a judgmental aid may be important to visual inferences (Skiba et al., 1989).

Morley and Adams (1991) define trend as "a systematic shift in the value of the central location of the data set over time" (p. 100). A quick impression of the presence of such a trend can be obtained by drawing a *vertical line graph*, in which the deviations from each data point to a measure of central tendency are plotted against time. An example of such a graph is given in the upper left corner of Figure 4. If a positive trend would be present in the data, the vertical lines on the left side would be hanging below the central tendency line, and the vertical lines on the right side would be standing on it. With the function `graph.TREND(design,TREND,CL,tr)` we included some possibilities to display a potential trend in the data. Depending on the desired central tendency measure, a vertical line graph can be drawn by setting the TREND argument to "`VLP`" and the `CL` argument to "`mean`," "`median`," "`bmed`," "`trimmean`," or "`mest`" (and if necessary complemented with the "`tr`" argument).

A trend in central location is usually displayed in single-case graphs by superimposing a linear function on the raw data, which shows if there is an increase or a decrease in the behavior over time. There is evidence that the use of such regression or trend lines can increase interrater agreement, reliability, and decision accuracy (Bailey, 1984; Fisher, Kelley, & Lomas, 2003; Hojem & Ottenbacher, 1988; Parsonson & Baer, 1992; Rojahn & Schulze, 1985; Skiba et al., 1989). However, results are not all positive, because in some cases trend lines created dependencies, helped maintaining inconsistent judgments, and led to an overemphasis of trend and the neglect of other factors like level and variability (DeProspero & Cohen, 1979; Fisch, 1998; Harbst et al., 1991; Hojem & Ottenbacher, 1988; Skiba, et al., 1989). Moreover, some researchers found no significant effect of using trend lines on the accuracy of decisions (Normand & Bailey, 2006). Fisch (1998) argues that these contradictory results could be partly explained by personal characteristics of the judges: experience seems to be more important than the use of visual guidelines. Trend lines are most easily drawn by 'eyeballing' a line that seems to bisect the distribution. Research has however shown that judges are not that good at deciding how and where exactly this line should be placed (Mosteller, Siegel, Trapido, & Youtz, 1981). Therefore we included some functions for fitting a robust linear trend through several methods.

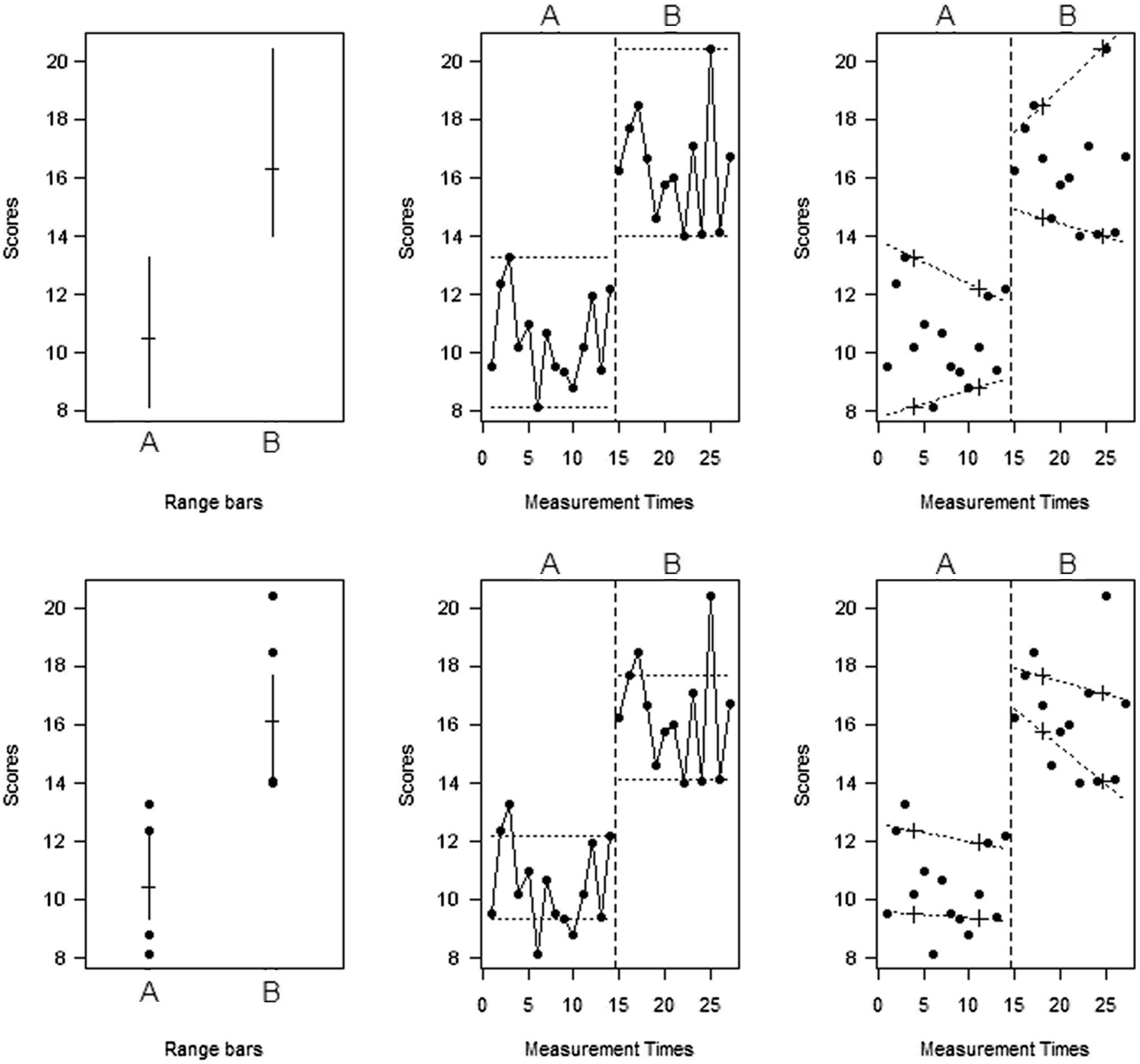


Figure 3. Hypothetical example of possible displays of variability for an AB phase design. The upper left panel shows range bars, with the mean as a measure of central tendency (`graph.VAR(design = "AB," VAR = "RB," CL = "mean")`), and the upper middle panel displays range lines (`graph.VAR(design = "AB," VAR = "RL")`). In the upper right panel, a trended range is shown (`graph. VAR(design = "AB," VAR = "TR")`). In the lower panels, the trimmed versions of each variability display are plotted: trimmed range bars on the left, trimmed range lines in the middle, and trimmed trended ranges on the right. This is done by adding the `dataset = "trimmed"` argument to the R command, for example, for range lines: `graph.VAR(design = "AB," VAR = "RL," dataset = "trimmed")`.

The first possibility to draw a line that approximates the best linear fit of the data consists of standard linear regression. When setting the TREND argument to "LSR" (least-squares regression), a regression line ($Y = a + bX$) that minimizes the squared vertical distances between the line and the data points is calculated from the slope and the intercept. An example is shown in the upper right corner of Figure 4. Just as with eyeballing, this method is however very much influenced by the presence of outliers. Several alternative, more robust, estimators have been proposed to replace ordinary least-squares regression. Wilcox (2005) gives a good overview, together with R functions to calculate those measures. The list of possibilities is very long (e.g., least median of squares, least trimmed squares, regression M-estimators with many variations, S-estimators), and we refer the interested reader to Wilcox (2005).

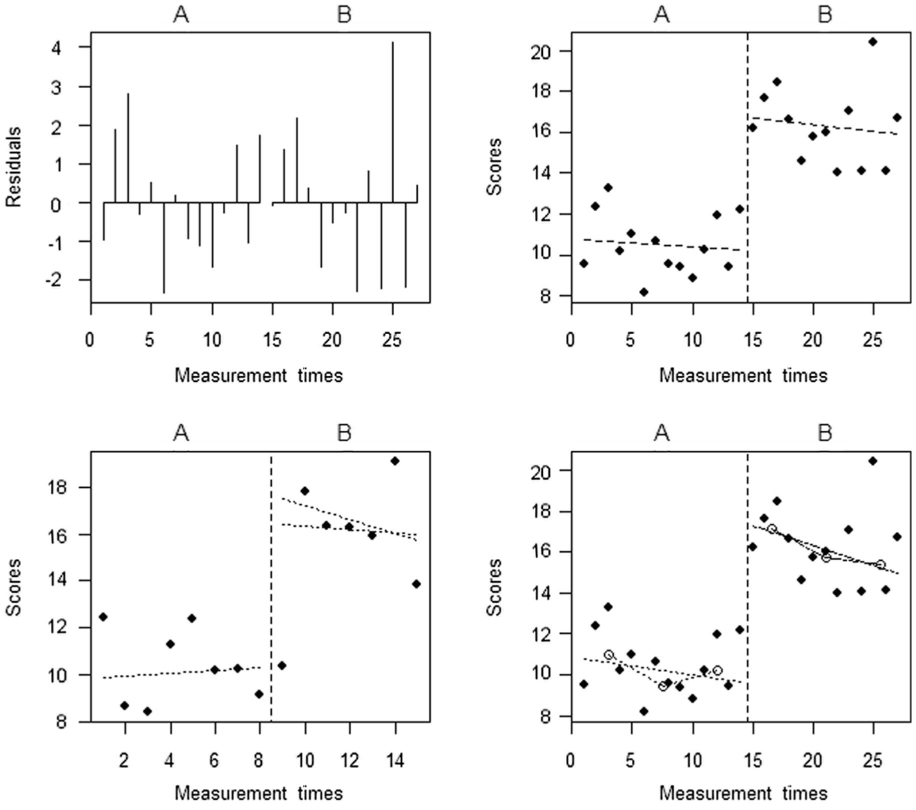


Figure 4. Hypothetical examples of linear trend displays for AB phase designs. In the upper left panel, a vertical line graph with deviations from the mean is displayed (graph.TREND(design = "AB," TREND = "VLP," CL = "mean")). The upper right panel shows trend lines by least-squares regression (graph.TREND(design = "AB,"TREND = "LSR")). Bottom left the split-middle method is demonstrated (graph.TREND(design = "AB," TREND = "SM")) and in the lower right corner we see the result of resistant trend line fitting (graph.TREND(design = "AB," TREND = "RTL")).

Another possibility to display trend is the *split-middle method* (TREND = "SM"), which is quite straightforward when phases have an even number of observations. The first step involves dividing the phase into two halves along the *X*-axis. When there are, for example, ten observations, the first five observations are allocated to the left phase half, and the last five observations are allocated to the right phase half. In the second step two values are calculated for each phase half: The median of the dependent variable and the middle time value. This middle time value is the observation in the middle of the time series for an uneven number of observations (e.g., the 3 in a series of five observations), whereas for an even number of observations this is the average of the two middle time values (e.g., 2.5 in a series of four observations), which is an imaginary time point. The final step of the split-middle method connects the crossing of the median (*Y*-axis) and the middle time value (*X*-axis) of both phase halves. The same three steps are followed when phases have an unequal number of observations, with as a difference that in the first step of dividing the phase into two halves, one of the halves will have an observation more than the other half. Therefore the division is conducted twice: once with the extra observation allocated to the left half and once with the extra observation allocated to the right half. Step two and step three are therefore also repeated, once for each way of allocating the data points, which results in two lines instead of one. The split-middle method is demonstrated in the lower left panel of Figure 4: in phase A the situation of an even number of observations is shown, while in phase B the two split-middle lines for a phase with an uneven number of observations are displayed.

Whereas the split-middle method is sufficient to display a linear trend for shorter time series (4–12 observations), the *resistant trend line fitting method* (TREND = "RTL") (Tukey, 1977) is more often used with larger time series. Both methods are comparable in finding the medians of the dependent variable and the middle time values, but with resistant trend line fitting the data are divided in three sections along the *X*-axis instead of two. The slope of the resistant trend line is determined by the change in the medians between the two outer data sections, and the intercept is computed using all three data sections to make the line pass as closely as possible through the middle of the data. By plotting the two half-lines, connecting the coordinates (intersection of the mid-time point and the median of the dependent variable) of each of the three sections, it is possible to check whether a linear fit is adequate. In our example, in the lower right panel of Figure 4, the fit for the B phase is almost linear, because the line connecting the coordinates of the two outer sections runs nearly through the midpoint of the middle section. This is not the case for the data in the A phase.

To further investigate the presence of a nonlinear trend in the data, we included the possibility of displaying *running medians*. With running medians, the time series is smoothed by dividing it into successive segments of a given size and calculating the median for each segment (Tukey, 1977). Three sizes of segments are easy to use with time-series data: running medians of three (`TREND = "RM3"`) can be considered when there are 6–10 observations (left panel of Figure 5), and for time series with more than 10 observations running medians of five (`TREND = "RM5"`) could be used (middle panel). These are calculated by finding the median of each batch of, respectively, 3 or 5 successive data points and plotting this value at the middle time point for that batch. This means that the first median value is plotted at the second time point (middle of 1-2-3) for RM3 and at the third time point (middle of 1-2-3-4-5) for RM5, reducing the time series by two (in the case of RM3) or four (in the case of RM5) data points. These two smoothers, however, often produce curves with high and low points. Another way of estimating the median of five successive values is by calculating running medians of batch size four, and then averaging each successive pair (`TREND = "RM42"`). As can be seen in the right panel of Figure 5 this results in a more smoothed curve. Of course many other smoothers exist, such as kernel and lowess smoothing. Wilcox (2005) provides R functions for making plots of these smoothed functions, as well as for more robust estimators.

From Graph to Data

To construct these graphical representations, one usually needs the original data values. This is also the case when statistically reanalyzing published data, for example to integrate the results from a study in a meta-analysis. Unfortunately, raw data are often not available from published studies. Therefore, we included a function to extract the values from existing graphs (adaptation of Timothée, 2010): `graph.extract(MT,refX,refY,save)`. Suppose that we want to do this for the data displayed in the left panel of Figure 5. We need to supply the number of observations (`MT = 15`), two reference values for the *X* scale (e.g., `refX=c(2,14)`), and two reference values for the *Y* scale (e.g., `refY=c(10,16)`). When giving the command, a window pops up in which the user can indicate where the jpeg file with the graph can be found. Then the reference values should be located on the graph by clicking with the left mouse button, first the *X* values and then the *Y* values, and the same should be done sequentially for each data point. As can be seen in the left panel of Figure 6, the marked reference points are indicated with a blue cross and the marked data points with a red circle. When finished, the calibrated true data points are displayed as output in the R console and plotted in a new graph. This is shown in the right panel of Figure 6. If one wants to save these data values to a file, the save argument should be put to "`yes`," and another window will pop up where it is asked in which file the data should be saved. Here one can create a new txt file in the folder of choice. Notice that for this function the R package ReadImages should be installed and the graphs of which one wants to extract data should be in jpeg format.

An overview of all graphical functions is given in Table 1. Table 2 summarizes the possible arguments.

Discussion

Although there have been serious concerns regarding the consistency, sensitivity, and specificity of visual analysis

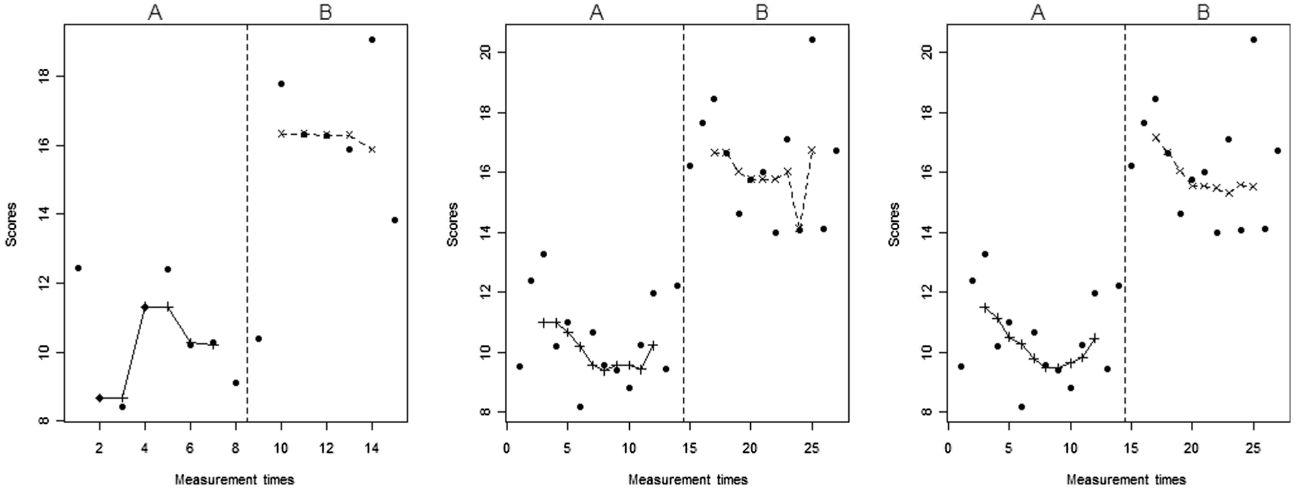
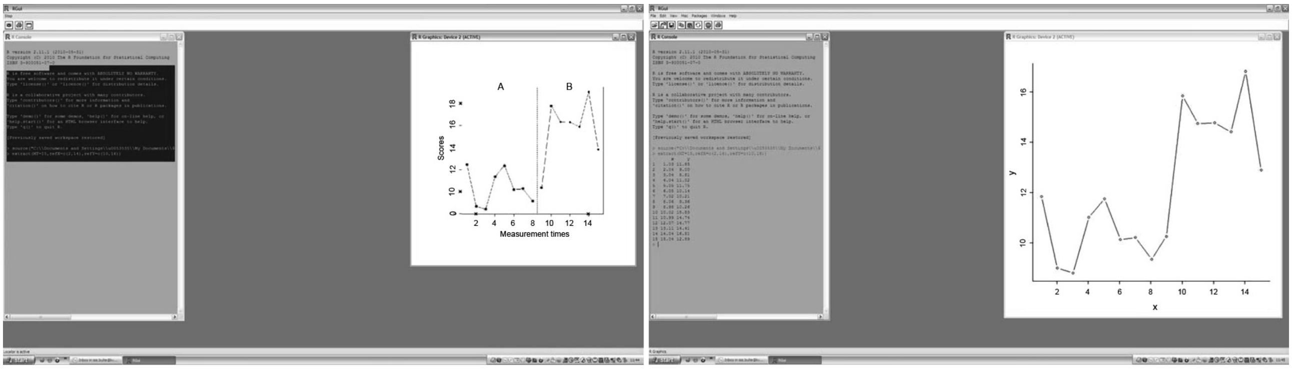


Figure 5. Examples of running medians superimposed on the hypothetical raw data. On the left, running medians of three are shown (`graph.TREND(design = "AB," TREND = "RM3")`). In the center we displayed running medians of five (`graph.TREND(design = "AB," TREND = "RM5")`), and on the right side running medians of four averaged by pairs are plotted (`graph.TREND(design = "AB," TREND = "RM42")`).



```
R version 2.11.1 (2010-05-31)
Copyright (C) 2010 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

> source("C:\\Documents and Settings\\u0053535\\My Documents\\
> extract(MT=15,refX=c(2,14),refY=c(10,16))
```

```
R version 2.11.1 (2010-05-31)
Copyright (C) 2010 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

> source("C:\\Documents and Settings\\u0053535\\My Documents\\
> extract(MT=15,refX=c(2,14),refY=c(10,16))
       x     y
1   1.03 11.85
2   2.04  9.00
3   3.04  8.81
4   4.04 11.02
5   5.05 11.75
6   6.05 10.14
7   7.02 10.21
8   8.06  9.36
9   8.98 10.26
10 10.02 15.85
11 10.99 14.74
12 12.07 14.77
13 13.11 14.41
14 14.04 16.81
15 15.04 12.89
> 
```

Figure 6. Print screen of the graph.extract function. On the left is shown how the reference values and data points are located on the graph. In the right panel the output is demonstrated: The calibrated true data points are returned in the R console and a new graph is created as a quick visual check.

of single-case data, this kind of analysis remains a necessary step in any thoughtful data-analytic procedure. In some instances, it could even be considered a sufficient step. When a visualization is that strong that the conclusions "hit you between the eyes" (what is sometimes called the "Intraocular Trauma Test"), statistical inference is of little relevance and *p* values do not really have a surplus value. When the effect of a behavior modification intervention is very clear, it is unnecessary to use complicated statistical modeling techniques such as time-series analysis to state the obvious. Graphs can also be sufficient when the purpose is to give feedback to individual patients. Graphical displays of single-case data are very well-suited for online monitoring of the patient's progress, by altering them every time new data become available, and for communicating results in an easy and understandable way. In other instances, of course, when the effect is not that plain to see or when the results have to be combined in a meta-analysis, statistical data analysis might be a useful supplementary technique. Even strong proponents of visual analysis acknowledge that statistical results can be valuable in the absence of a stable baseline, when results must be shared unambiguously with other professionals, and when testing new treatments of which the effects cannot be predicted (Parker et al., 2005).

This possible need for complementary statistical analyses can be easily accomplished by working in R. We did not only choose this programming language because of its graphical possibilities, but also because it is a statistical environment, which facilitates the combined use of visual and statistical analyses. Moreover, as indicated before, the flexibility of R enables users to change parts of the code according to their demands (e.g., labels of the axes). This requires some experience with the program, but this is easily acquired. We acknowledge that there might be a threshold to start working with R, because of the lower user-friendliness of its command line interface. However, to use the functions described in this paper in their standard versions, one does not need to know much about R. After downloading and installing R (Hornik, 2010, gives a detailed explanation of how to do this for Windows, Macintosh, and UNIX), the functions can be read into R by choosing "File" and selecting "Source R Code" from the top menu. For easy access and use, we suggest saving the files containing the R functions (which can be found on ppw.kuleuven.be/english/research/mesrg) on the local disk. The text files (.txt) containing the data are most easily made in a text editor (e.g., EditPad or NotePad) or in Excel (saved as "text (tab delimited)"). For phase designs and alternating treatments designs the file should consist of two columns: one containing the condition labels ("A" and "B") and the second with the observations. It is important not to label the rows or columns. For multiple baseline AB designs the data file should consist of these two columns for each unit (so twice as many columns as there are units). In the future, more and more forms of graphical user interfaces for R will become available, what will enhance the user-friendliness (but maybe also decrease the flexibility).

Table 1. Overview of graphical functions

Function name	Description
`graph(design)`	Makes a graphical representation of the single-case data
`graph.CL(design,CL,tr)`	Plots a measure of central tendency as a horizontal reference line superimposed on the raw time-series data
`graph.VAR(design,VAR,dataset,CL,tr)`	Displays information about variability in the data
`graph.TREND(design,TREND,CL,tr)`	Visualizes systematic shifts in central location of a data set over time
`graph.extract(MT,refX,refY,save)`	Extracts data values from published graphs

Table 2. Overview of graphical arguments

Argument	Values
`design`	Indicates what type of single-case design has been used: A phase design (“`AB`,” “`ABA`,” “`ABAB`”), an alternating treatments design (“`ATD`”), or a multiple-baseline AB design (“`MBD`”)
`CL`	Indicates what measure of central tendency should be plotted: “`mean`,” “`median`,” “`bmed`” (broadened median), “`trimmean`” (trimmed mean), or “mest” (M-estimator)
`tr`	For a trimmed mean: indicates which proportion (0–0.5) of the observations has to be removed from each side before calculating the mean. A common choice when calculating a trimmed mean is tr = 0.2 For a M-estimator of location: indicates the constant against which the observations are evaluated to determine if they are extreme. This constant is usually chosen at 1.28 (the 0.9 quantile of the standard normal distribution) (Wilcox, 2005)
`VAR`	Indicates what kind of variability information should be plotted: range lines (“`RL`”), range bars (“`RB`”), or a trended range (“`TR`”)
`dataset`	Indicates whether the whole data set should be used (“`regular`” = default) or the 10–20% extreme values from each phase should be removed (“`trimmed`”)
`TREND`	Indicates what trend visualization should be plotted: vertical line plot (“`VLP`”); trend lines by means of least-squares regression (“`LSR`”), split-middle method (“`SM`”), resistant trend line fitting (“`RTL`”), or running medians (“`RM3`,” ”`RM5`,” “`RM42`”)
`MT`	Number of observations (measurement times)
`refX`	Reference values for the *X* scale: `c(x1,x2)`
`refY`	Reference values for the *Y* scale: `c(y1,y2)`
`save`	Indicates whether the data values should be saved to a file (`save =` “`yes`”) or just displayed as output in the R console (`save =` “`no`” = default)

Acknowledgment

This research was funded by the Research Foundation – Flanders (FWO).

References

Allison, D. B., Franklin, R. D., & Heshka, S. (1992). Reflections on visual inspection, response guided experimentation, and Type I error rate in single-case designs. *The Journal of Experimental Education, 61*, 45–51.

Bailey, D. B. Jr. (1984). Effects of lines of progress and semilogarithmic charts on ratings of charted data. *Journal of Applied Behavior Analysis, 17*, 359–365.

Barlow, D. H., Nock, M. K., & Hersen, M. (2009). *Single case experimental designs: Strategies for studying behavior change* (3rd ed.). Boston, MA: Allyn & Bacon.

Bobrovitz, C. D., & Ottenbacher, K. J. (1998). Comparison of visual inspection and statistical analysis of single-subject data in rehabilitation research. *American Journal of Physical Medicine & Rehabilitation, 77*, 94–102.

Brossart, D. F., Parker, R. I., Olson, E. A., & Lakshmi, M. (2006). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. *Behavior Modification, 30*, 531–563.

Busk, P. L., & Marascuilo, L. A. (1992). Statistical analysis in single-case research: Issues, procedures, and recommendations, with special applications to multiple behaviors. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research designs and analysis: New directions for psychology and education* (pp. 159–185). Hillsdale, NJ: Erlbaum.

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). New York, NY: Routledge.

Carter, M. (2009). Effects of graphing conventions and response options on interpretation of small n graphs. *Educational Psychology, 29*, 643–658.

DeProspero, A., & Cohen, S. (1979). Inconsistent visual analyses of intrasubject data. *Journal of Applied Behavior Analysis, 12*, 573–579.

Ferron, J., & Jones, P. K. (2006). Tests for the visual analysis of response-guided multiple-baseline data. *The Journal of Experimental Education, 75*, 66–81.

Fisch, G. S. (1998). Visual inspection of data revisited: Do the eyes still have it? *The Behavior Analyst, 21*, 111–123.

Fisher, W. W., Kelley, M. E., & Lomas, J. E. (2003). Visual aids and structured criteria for improving visual inspection and interpretation of single-case designs. *Journal of Applied Behavior Analysis, 36*, 387–406.

Franklin, R. D., Gorman, B. S., Beasley, T. M., & Allison, D. B. (1997). Graphical display and visual analysis. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 119–158). Mahwah, NJ: Erlbaum.

Furlong, M. J., & Wampold, B. E. (1982). Intervention effects and relative variation as dimensions in experts' use of visual inference. *Journal of Applied Behavior Analysis, 15*, 415–421.

Gibson, G., & Ottenbacher, K. (1988). Characteristics influencing the visual analysis of single-subject data: An empirical analysis. *The Journal of Applied Behavioral Science, 24*, 298–314.

Gorman, B. S., & Allison, D. B. (1997). Statistical alternatives for single-case designs. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 159–214). Mahwah, NJ: Erlbaum.

Harbst, K. B., Ottenbacher, K. J., & Harris, R. S. (1991). Interrater reliability of therapists' judgments of graphed data. *Physical Therapy, 71*, 107–115.

Hojem, M. A., & Ottenbacher, K. J. (1988). Empirical investigation of visual-inspection versus trend-line analysis of single-subject data. *Physical Therapy, 68*, 983–988.

Hornik, K. (2010). *The R FAQ.* Retrieved from CRAN.R-project.org/doc/FAQ/

Houle, T. T. (2009). Statistical analyses for single-case experimental designs. In D. H. Barlow, M. T. Nock, & M. Hersen (Eds.), *Single case experimental designs: Strategies for studying behavior change* (3rd ed., pp. 271–305). Boston, MA: Allyn & Bacon.

Huber, P., & Ronchetti, E. (2009). *Robust statistics* (2nd ed.). Hoboken, NJ: Wiley.

Jeng, H. (2010). On small samples and the use of robust estimators in loss reserving. *Casualty Actuarial Society E-Forum, 1*, 1–27.

Jones, R. J., Weinrott, M. R., & Vaught, R. S. (1978). Effects of serial dependency on the agreement between visual and statistical inference. *Journal of Applied Behavior Analysis, 11*, 277–283.

Kahng, S. W., Chung, K.-M., Gutshall, K., Pitts, S. C., Girolami, J. K., & Girolami, K. (2010). Consistent visual analyses of intrasubject data. *Journal of Applied Behavior Analysis, 43*, 35–45.

Kazdin, A. E. (2011). *Single-case research designs: Methods for clinical and applied settings*. New York, NY: Oxford University Press.

Kelley, K. (2007). Methods for the behavioral, educational, and social sciences: An R package. *Behavior Research Methods, 39*, 979–984.

Knapp, T. J. (1983). Behavior analysts' visual appraisal of behavior change in graphic display. *Behavioral Assessment, 5*, 155–164.

Kratochwill, T. R., & Brody, G. H. (1978). Single subject designs: A perspective on the controversy over employing statistical inference and implications for research and training in behavior modification. *Behavior Modification, 2*, 291–307.

Long, C. G., & Hollin, C. R. (1995). Single case design: A critique of methodology and analysis of recent trends. *Clinical Psychology & Psychotherapy, 2*, 177–191.

Ma, H. H. (2006). An alternative method for quantitative synthesis of single-subject research: Percentage of data points exceeding the median. *Behavior Modification, 30*, 598–617.

Matyas, T. A., & Greenwood, K. M. (1990). Visual analysis of single-case time series: Effects of variability, serial dependence and magnitude of intervention effects. *Journal of Applied Behavior Analysis, 23*, 341–351.

Morley, S., & Adams, M. (1991). Graphical analysis of single-case time series data. *British Journal of Clinical Psychology, 30*, 97–115.

Mosteller, F., Siegel, A. F., Trapido, E., & Youtz, C. (1981). Eye fitting straight lines. *The American Statistician, 35*, 150–152.

Normand, M. P., & Bailey, J. S. (2006). The effects of celeration lines on visual data analysis. *Behavior Modification, 30*, 295–314.

Ottenbacher, K. J. (1990). When is a picture worth a thousand *p* values? A comparison of visual and quantitative methods to analyze single subject data. *Journal of Special Education, 23*, 436–449.

Park, H.-S., Marascuilo, L., & Gaylord-Ross, R. (1990). Visual inspection and statistical analysis in single-case designs. *The Journal of Experimental Education, 58*, 311–320.

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*, 116–132.

Parker, R. I., Hagan-Burke, S., & Vannest, K. (2007). Percentage of all non-overlapping data: An alternative to PND. *Journal of Special Education, 40*, 194–204.

Parsonson, B. S., & Baer, D. M. (1992). The visual analysis of data, and current research into the stimuli controlling it. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 15–40). Hillsdale, NJ: Erlbaum.

Richards, S. B., Taylor, R. L., & Ramasamy, R. (1997). Effects of subject and rater characteristics on the accuracy of visual analysis of single subject data. *Psychology in the Schools, 34*, 355–362.

Rojahn, J., & Schulze, H. H. (1985). The linear regression line as a judgmental aid in visual analysis of serially dependent A-B time-series data. *Journal of Psychopathology and Behavioral Assessment, 7*, 191–206.

Skiba, R., Deno, S., Marston, D., & Casey, A. (1989). Influence of trend estimation and subject familiarity on practitioners judgments of intervention effectiveness. *Journal of Special Education, 22*, 433–446.

Timothée. (2010, March 5). Data visualization (in R). Getting data from an image (introductory post) [Web log message]. Retrieved from http://rdataviz.wordpress.com/2010/03/05/getting-data-from-an-image-introductory-post/

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Wampold, B. E., & Furlong, M. J. (1981). The heuristics of visual inference. *Behavioral Assessment, 3*, 79–92.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd ed.). San Diego, CA: Elsevier.

Wilkinson, L., the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594–604.

Wilkinson, L. (2005). *The grammar of graphics* (2nd ed.). New York, NY: Springer.

Ximenes, V. M., Manolov, R., Solanas, A., & Quera, V. (2009). Factors affecting visual inference in single-case designs. *The Spanish Journal of Psychology, 12*, 823–832.

Received September 1, 2010
Accepted February 21, 2011
Published online August 5, 2011

Isis Bulté

Methodology of Educational Sciences Research Group
Andreas Vesaliusstraat 2
Box 3762
B-3000 Leuven
Belgium
Tel. +32 16 326201
Fax +32 16 326200
E-mail isis.bulte@ped.kuleuven.be