




Graphing Within-Subjects Effects

David M. Lane ^a, 

^aDepartments of Psychological Science, Statistics, and Management, Rice University

Abstract ■ Most graphs in psychology articles fail to show distributional information other than the mean. Although many articles have suggested solutions to this problem for between-subjects designs, there has been relatively little discussion on how to show distributional information for within-subjects designs. Graphs of within-subjects data should be constructed so that between-subjects variation does not appear as uncontrolled error. This article presents a variety of methods for graphing data from within-subjects designs including jittered dot plots of difference scores, sum-difference plots, box plots of difference scores, and plots of trend components. These graph types show descriptive distributional information while controlling for between-subjects variation.

Keywords ■ Graphing data; within-subject data; jittered plots, sum-different plots; box plots, plots of trends..

 lane@rice.edu

 DML: 0000-0002-6364-9945

 [10.20982/tqmp.15.3.p174](https://doi.org/10.20982/tqmp.15.3.p174)

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Reviewers

■ Thom Baguley (Nottingham Trent University)

■ Fernando Marmolejo-Ramos (Adelaide University)

■ and one more.

Introduction

Wilkinson and the APA Task Force on Statistical Inference (1999) argued that a common deficiency of graphs is their failure to display distributional information. Criticizing the use of graphs such as bar charts that do not portray distributional information, they argued that other graph types such as box plots and stem and leaf plots are preferable and should be used instead. They further argued that:

A common deficiency of graphics in psychological publications is their lack of essential information. In most cases, this information is the shape or distribution of the data. Whether from a negative motivation to conceal irregularities or from a positive belief that less is more, omitting shape information from graphics often hinders scientific evaluation. (p. 601)

Numerous more recent articles concur with Wilkinson and the APA Task Force on Statistical Inference (1999) on this (Correll & Gleicher, 2014; Duke et al., 2015; Krzywinski & Altman, 2014; Lane & Sandor, 2009; Larsen-Hall, 2017; Martinez, 2015; Marmolejo-Ramos & Matsunaga, 2009; Weissgerber, Milic, Winham, & Garovic, 2015).

Graphing distributional information from within-subjects designs faces a difficulty not found for between-

subjects designs: in within-subjects designs, variance due to differences between subjects is controlled and therefore should not be displayed as random variation. This is similar to the problem of displaying confidence intervals for data from within-subjects designs for which between-subjects variation should not affect the confidence intervals. Methods for constructing confidence intervals for within-subjects designs that do not confound between- and within-subjects variations have been presented by Baguley (2012), Cousineau and O'Brien (2014), and Loftus and Masson (1994).

Graphs can portray distributions and/or differences between distributions as well as relevant inferential statistics such as confidence intervals and significance tests. Although methods for including inferential statistics in graphs is an important topic and has been addressed in detail (Cousineau & O'Brien, 2014; Cumming & Finch, 2005; Wright, Klein, & Wieczorek, 2019), it is beyond the scope of this article to discuss the pros and cons of the various approaches or to recommend a specific one. Therefore, the example graphs shown here do not contain any inferential statistics. This should not be taken as a suggestion that authors should not supplement the kinds of graphs discussed here with inferential statistics in the manner they deem most appropriate.



This article uses four datasets to illustrate ways of displaying data for a variety of designs. Three of the datasets are from real experiments whereas the fourth contains fictitious data created to facilitate the illustration of graphical methods. The example graphs all portray distributions and/or differences between distributions in ways that control for between-subjects variance.

The first dataset is used to illustrate ways to display (a) paired data with and without a between-subjects variable and (b) linear contrasts among conditions. The second gives two examples of how to display data from a one-way within-subjects design with more than two levels. One example shows a method for graphing differences in successive levels of a variable and the other shows how the number of graphs necessary to accurately reflect the data can be reduced if the assumption of sphericity is at least approximately met. The third dataset is used to illustrate how to display data from a design with a between-subjects variable and a within-subject variable when the assumption of sphericity is not met. The fourth dataset illustrates how to graph the results of a trend analysis in a design with a between-subjects variable and a within-subjects variable.

Anderson, Benjamin, and Bartholow (1998): Illustrating graphs for paired data and for linear contrasts

This section contains examples of displaying (a) paired data with and without a between-subjects variable and (b) linear contrasts among conditions. The data for this study are from an experiment by Anderson et al. (1998). Their experimental design consisted of one between-subjects variable: (gender) and two within-subjects variables: type of priming word (weapon or non-weapon) and type of target word (aggressive or non-aggressive control). The dependent variable was the time to name the target word. There were 32 subjects in the experiment.

Paired data, no between-subjects variables

An excellent way to portray paired data is to connect the points of the paired observations creating a graph type called a “comparative graph” (Harris, 1999), a “parallel coordinate graph” (Schriger, 2017), or a strip chart with lines in R documentation (Millard, 2013). The lines connecting the two points for each observation are important because they show the consistency or inconsistency of differences across subjects. Figure 1a is a comparative graph showing times for aggressive words primed by non-weapon words (NA) and primed by weapon words (WA) from a randomly-selected subset of 14 out of the 32 subjects. Figure 1a shows that the times for the WA condition were consistently although not uniformly shorter than the times for the NA

condition. In addition to what could be learned from a graph of means, Figure 1a show that the effect is very small for most subjects, fairly large for two, moderately large for three, and strongly in the opposite direction for one. It also shows that there are no outliers in overall response time or in the effect of conditions.

Much less informative is available in Figure 1b that shows the same data without the lines connecting the paired observations. Without the lines, the large differences among subjects obscure the consistent difference between conditions. The report of the significant effect of conditions, $t(13) = 2.64, p = 0.020$, in conjunction with a graph such as Figure 1b would likely seem incongruous to many readers.

Comparative graphs are most effective when there are relatively few observations (Harris, 1999; McNeil, 1992) although Harris (1999) noted that comparative graphs can be effective even with a large number of observations when the emphasis is on patterns and trends. Graphs showing difference scores can more easily accommodate large numbers of observations than can comparative graphs and, therefore, can be a good choice when comparative graphs are not effective. Among the possible ways to graph difference scores are box plots, histograms, stem and leaf displays, jittered dot plots, and density plots. Figure 2 shows a jittered dot plot of differences between the NA and WA conditions for all 32 subjects in the experiment. Unlike a graph of means, Figure 2 shows that although the differences were primarily positive (NA higher than WA), the effect was in the opposite direction for about one third of the subjects, the data are not highly skewed, and that there are no outliers.

Although plots of difference scores reveal all the information relevant to the usual significance test, other information of possible theoretical importance is hidden. Sum-difference graphs¹ proposed by John Tukey (described in Cleveland, 1994) show more information. In these graphs, differences between pairs are shown on the Y-axis, their sums on the X-axis, and a horizontal line is drawn at zero. For example, Figure 3a shows sum-difference graphs for the data from Anderson et al. (1998). This figure shows that the differences between NA and WA tend to be larger for both low and high values of the sum. This is consistent with the results of a polynomial regression finding a significant quadratic term, $t(29) = 2.70, p = 0.012$. Gender differences are considered in the next section.

A different pattern is apparent in Figure 3b which shows fictitious data created so that the difference scores are identical to those in Figure 3a but observations with generally larger scores (higher sums) have larger differ-

¹Sum-difference plots are equivalent to a Bland-Altman analysis (Altman & Bland, 1983) although the latter were originally created for a different purpose.

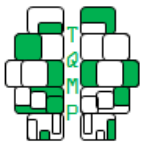
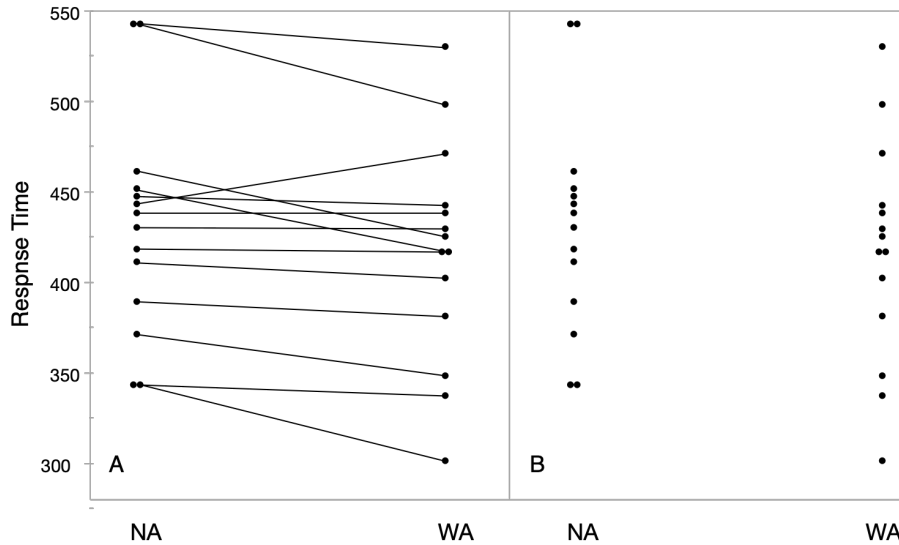


Figure 1 ■ Comparative graph (left) and jittered dot plot (right).



ences. The correlation between sums and differences is 0.49 and the slope is 0.114, $p = 0.004$. The key point is that two sum-difference graphs can show very different patterns even when graphs of difference scores are identical.

Paired data with a between-subjects variables

Many experimental designs contain both between-subjects and within-subjects variables. The effect of groups, the effect of trials, and the Groups \times Trials interaction are typically of interest in these designs.

Whenever there is a non-trivial difference between groups, it is important for graphs of a within-subjects effect to show the data for the groups separately. Otherwise, differences between groups will increase the apparent variability of the difference scores. For example, this would occur if the difference scores were much larger in one group than another.

There are many ways to graph differences between groups. Although bar charts are very common, they are a poor choice because they reveal no descriptive distributional information other than the means themselves. Better alternatives include box plots, back-to-back stem and leaf plots, and dot plots. Figure 4 contains box plots of differences between the NA and WA conditions as a function of gender. No evidence of a Gender \times Condition interaction (gender difference in the difference between NA and WA) is apparent. Further, the distributions do not appear to be highly skewed and there are no outliers. Although there are many varieties of box plots, those that show the means

have the advantage of being consistent with the inferential statistics when the inferential statistics reported are tests of means.

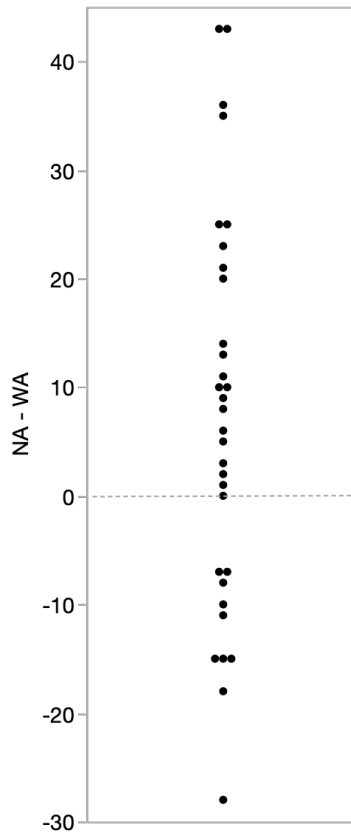
Sum-difference graphs can be designed to include a between-subjects variable. For example, Figure 3a uses filled rectangles for females and unfilled rectangles for males. Although the slope of the relationship between the sum and the difference scores is steeper for the females (-0.062) than for the males (-0.005), neither the slope for females, $t(13) = -0.99, p = 0.340$, the slope for males, $t(5) = -0.12, p = 0.908$, nor the difference in slopes is significant, $F(1, 28) = 0.50, p = 0.484$. Although the mean for the males (8.18) is slightly higher than the mean for the females (6.00), this difference is not significant either, $F(1, 30) = 0.11, p = 0.742$. Finally, although the quadratic component is stronger for males than for females, the Gender \times Quadratic component is not significant, $F(1, 26) = 1.76, p = 0.196$. These inferential statistics show the importance of using inferential statistics to keep from over interpreting patterns observed in graphs.

Linear contrasts

Comparisons more complex than simple pairwise differences are often required to investigate the primary research question. The study by Anderson et al. included a control-word condition with the critical question being whether the difference in priming between a weapon-word prime and a non-weapon-word prime is greater when the primed word is an aggressive word than when it is a non-aggressive control word. Defining WC and NC as a



Figure 2 ■ Jittered point plot of difference scores for all 32 subjects in the experiment.



control word primed by a weapon word and a control word primed by a non-weapon word respectively, the critical question is whether $(NA - WA) - (NC - WC)$ is greater than zero. This difference between differences is an interaction contrast because it compares the size of the priming effect on aggressive words with the size of the priming effect on control words. The sum-difference plot in Figure 5 portrays the distribution of this contrast and shows that scores are predominantly positive with the larger values of the contrast occurring for small and for large values of the sum. As with the difference between NA and WA, the quadratic component of the relationship is significant, $t(29) = 2.58, p = 0.015$. In this graph, the size of the interaction contrast is plotted as a function of total response time. Alternatively, it is possible to plot the size of an interaction contrast as a function of one or more of the main effects. For example, the main effect of prime type could be represented on the X-axis as $NA + NC - WA - WC$.

Pearson et al. (2004): Graphs of successive differences and when sphericity is met

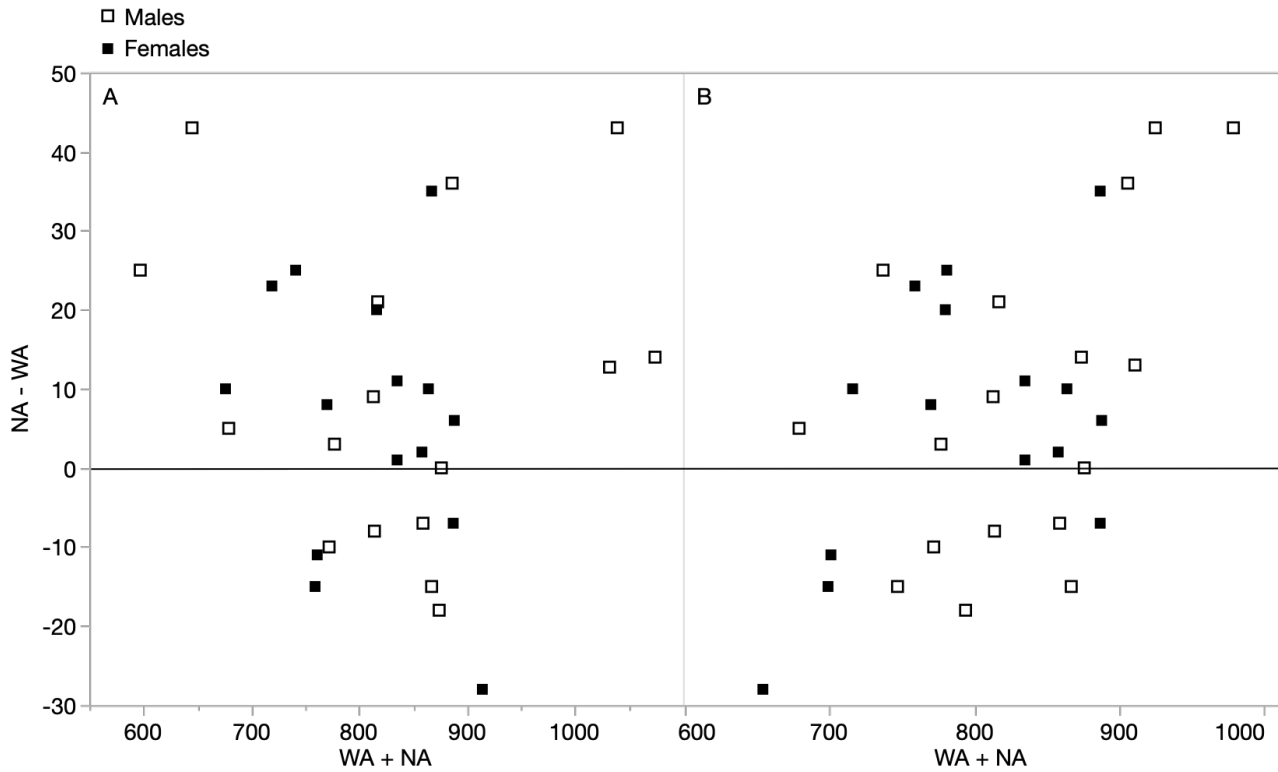
This section gives examples of graphs from a one-way within-subjects design with four levels. The first example displays differences between successive dosage levels whereas the second displays all pairwise differences using a method that assumes that the variances of all pairwise differences are equal, an assumption known as “sphericity.” The data are from a study of the effect of methylphenidate on the performance of 24 children diagnosed with ADHD (Pearson et al., 2004). Each child’s performance on a delay of gratification task was measured for each of the dosage levels 0 mg, 15 mg, 30 mg, and 60 mg.

Successive dosages

It is natural to consider performance differences as a function of successive dose increases. Figure 6 contains box plots of successive differences between the 0 mg, 15 mg, 30 mg, and 60 mg conditions. These graphs are consistent with the inferential statistics showing that only the D30-



Figure 3 ■ Sum and difference plots of the NA and WA conditions. Figure 3a (left) contains the actual data from the experiment whereas Figure 3b contains modified data with the difference scores unchanged.



D15 difference is significant, two-tailed $p = 0.006$. Further, they show that there may be some positive skew in the D15-D0 difference and that there is one outside value in the D60-D30 difference.

Assumption of sphericity met

Graphs of all pairwise comparisons would have resulted in six box plots thus making it somewhat difficult to comprehend all the results quickly. Alternatively, one graph can be created to accurately represent all four condition by adjusting the data to control between-subjects differences. As noted by others (Bakeman & McArthur, 1996; Loftus, 1995), between-subjects differences can be controlled by adjusting each of a subject's scores for the overall level of performance of that subject. The steps are:

1. Compute the mean across conditions separately for each subject.
2. Subtract each subject's mean from each of their scores.
3. Add the mean for each condition to each subject's score in that condition.

For example, if the four scores for a subject were 4, 6, 9, and 13, then the mean for the subject would be 8 and the

deviations from the subject means would be -4, -2, 1, and 5. If the mean of all subjects in this subject's condition were 6, then adding this mean to the scores would result in adjusted scores of 2, 4, 7, and 11. Since the effect of subjects is controlled, the sum of squares for subjects would be 0 in a model predicting these adjusted scores based on condition and subject. The F for condition would match the F in a standard repeated measures ANOVA.

Figure 7 contains box plots of the four conditions of the Pearson et al. (2004) study for the raw scores and the adjusted scores. This figure shows that the means for the raw and adjusted scores are the same whereas the variability of the adjusted scores is considerably reduced. Therefore, effect sizes measured in terms of mean differences relative to variability are increased by the adjustment. Figure 7 also shows that the variability of the adjusted scores in the D0 condition is considerably lower than in the other conditions except for the presence of two outside values. Finally, Figure 7 shows that the adjusted scores in the D60 condition are negatively skewed.

The use of adjusted scores does not accurately portray the variability of differences between conditions unless the

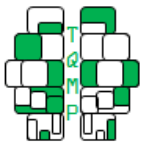
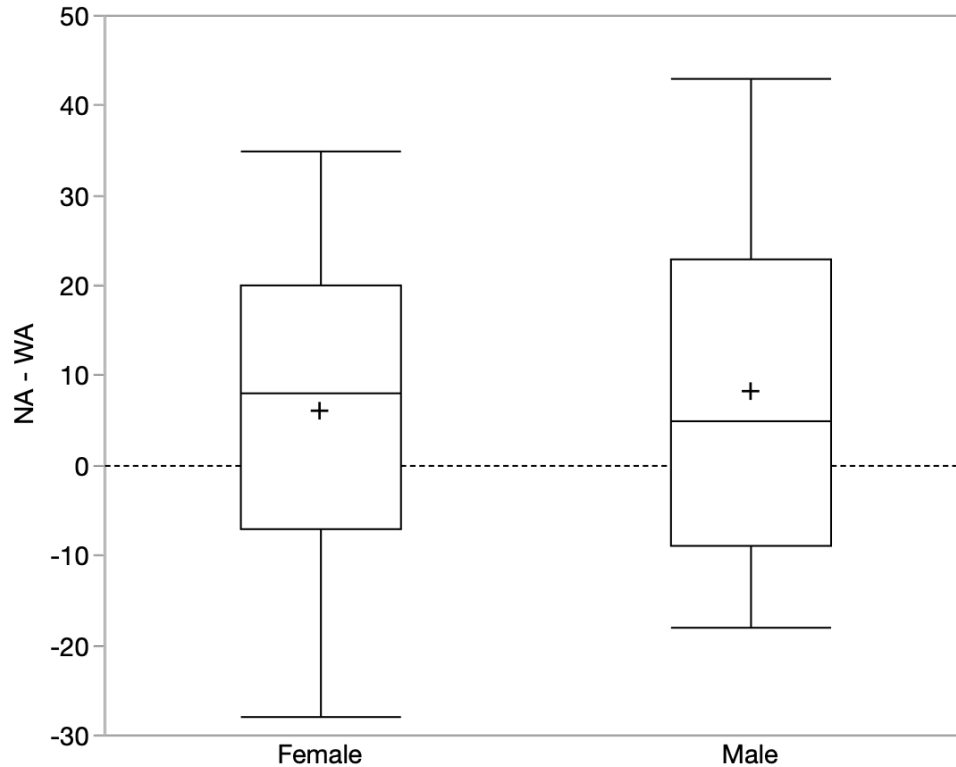


Figure 4 ■ Box plots of difference scores. The "+" signs represent the means.



assumption of sphericity that the variances of all pairwise differences are equal is at least approximately met. The variances of difference scores for Pearson et al.'s data are 65.91, 67.21, 56.82, 56.06, 94.39, and 90.33 which are similar and, as a result, the deviation from sphericity is not significant, $\chi^2(5) = 4.05, p = 0.543$. Therefore, the adjusted box plots in Figure 7 adequately represent the variability of differences between conditions.

In-class experiment illustrating graphs when the sphericity assumption is not met

This example presents a graph based on data for which the assumption of sphericity that all pairwise differences among the levels of a within-subjects variable are equal is violated. The data are from a Stroop experiment conducted in an undergraduate statistics class in which 47 students named a set of colored rectangles (colors), color names (words), and the ink color of colored words for which the ink color and the color name conflicted (interference).

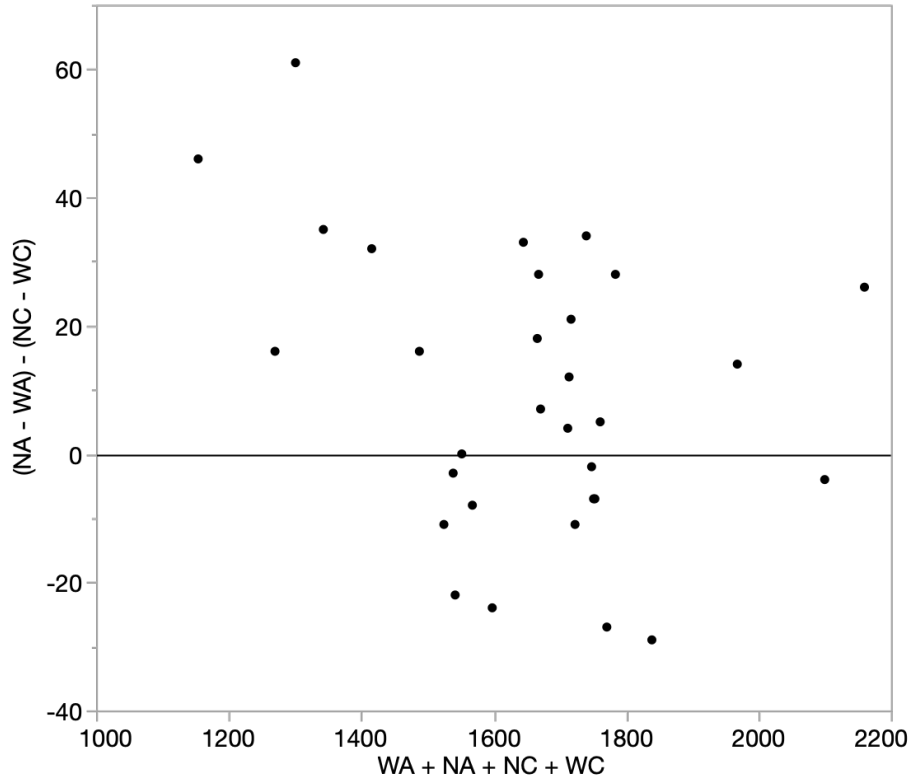
The variances of differences between times for colors

and words, colors and interference, and words and interference were 8.91, 55.80, and 61.52 seconds squared respectively thus greatly violating the assumption of sphericity. Adjusted box plots would be misleading because they would not reveal these differences in variability. Therefore, for these data it is better to display all three box plots of pairwise difference scores as is done in Figure 8. Since there is a between-subjects variable (gender), there are two box plots for each pairwise difference. The differences in variability noted above are clear from Figure 8. Also notable is that the distribution of interference minus colors shows a positive skew whereas there is little if any skew in the other distributions. Finally, there is an outside value for the females in the interference minus colors condition.

This approach of plotting all pairwise differences can be problematic when there are many levels of the within-subjects variable. For example, there are 45 comparisons among 10 levels. In general, there are $m(m-1)/2$ pairwise comparisons if there are m levels of the within-subjects variable. One approach in these cases is to select a subset of comparisons that are of particular importance and



Figure 5 ■ Sum and difference plot for the contrast (NA-WA)-(NC-WC).



only graph distributional data for those.

Fictitious Data Illustrating Graphs of Components of Trend

In repeated measures and longitudinal designs, correlations between trials occurring closer in time tend to be more highly correlated than those farther apart, a pattern referred to as a “simplex configuration” (Wallenstein & Fleiss, 1979). Since higher correlations between trials are associated with lower variances of difference scores, the simplex configuration results in unequal variances of difference scores and thus violates sphericity. The implica-

tions of this are explored in the context of a hypothetical learning experiment with two treatment groups (experimental and control) and six trials.²

Figure 9 shows a line graph of the means over trials for the two conditions. Since adding descriptive distributional information to this graph would at least partially obscure the pattern, descriptive distributional information is shown in a separate graph.

Figure 9 reveals that performance increases over trials for both groups and that the increase is greater for the experimental group. Further, the function is close to linear for the experimental group and negatively accelerated

²The data were generated by sampling randomly from a normally-distributed population with the following covariance matrix.

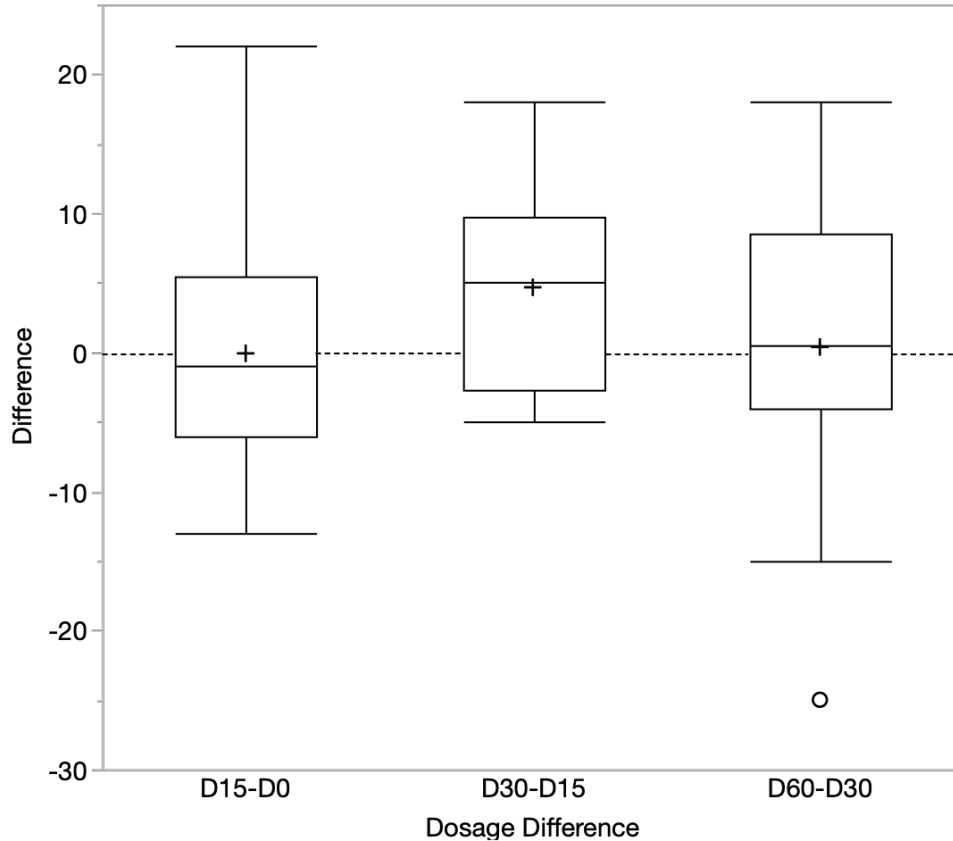
$$\begin{matrix}
 10 & 5 & 4 & 3 & 2 & 1 \\
 5 & 10 & 5 & 4 & 3 & 2 \\
 4 & 5 & 10 & 5 & 4 & 3 \\
 3 & 4 & 5 & 10 & 5 & 4 \\
 2 & 3 & 4 & 5 & 10 & 5 \\
 1 & 2 & 3 & 4 & 5 & 10
 \end{matrix} \tag{1}$$

and the following means

$$\begin{matrix}
 \text{Experimental :} & 10 & 11, & 12, & 13, & 14 & 15 \\
 \text{Control :} & 10, & 11, & 12, & 12, & 13, & 13
 \end{matrix} \tag{2}$$



Figure 6 ■ Box plots of successive dosage differences, The "+" signs represent the means.



for the control group. Assume that, as is often the case in learning experiments, the researchers were, a priori, interested in the linear and quadratic components of trend but had little interest in higher-order trend components.

A test of trend components shows that both the Group \times Trials (linear), $F(1, 58) = 5.79, p = 0.019$, and the Group \times Trials (quadratic), $F(1, 58) = 8.28, p = 0.006$, interactions are significant. The linear component is significant for the the Control Group, $t(29) = 4.27, p < 0.001$, and the Experimental Group, $t(29) = 8.37, p < 0.001$. The quadratic component is significant for the Control Group, $t(29) = 4.96, p < 0.001$ but not for the Experimental Group, $t(29) = -0.19, p = .849$. Thus, there is strong evidence that the linear component is stronger for the Experimental Group than for the Control Group while the reverse is true for the quadratic component. There is strong evidence of a linear component for both groups but evidence

of a quadratic component only for the Control Group.

Displaying box plots for all 15 pairwise differences would fail to bring the linear and quadratic components of trend into focus whereas displaying adjusted box plots would be misleading because of the violation of sphericity. The alternative shown in Figure 10 involves computing the linear and quadratic components separately for each subject and constructing box plots comparing the conditions for each component.

Figure 10 was constructed first by normalizing³ the trend coefficients for the linear and quadratic trends so that the sum of squared coefficients for each set of coefficients is 1. The normalized coefficients are $-0.60, -0.36, -0.12, 0.12, 0.36$, and 0.60 for the linear trend and $0.55, -0.11, -0.44, -0.44, -0.11$, and 0.55 for the quadratic trend. The next step was to create new variables by applying these coefficients to the raw data as fol-

³Normalization is achieved by (a) summing the squared coefficients, (b) taking the square root of the sum, and (c) dividing each coefficient by this square root. For the linear trend coefficients, this is done by dividing each of the typical linear trend coefficients $(-5 \ -3 \ -1 \ 1 \ 3 \ 5)$ by the square root of $(25 + 9 + 1 + 1 + 9 + 25)$ to get $-0.60, -0.36, -0.12, 0.12, 0.36, 0.60$.

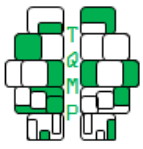
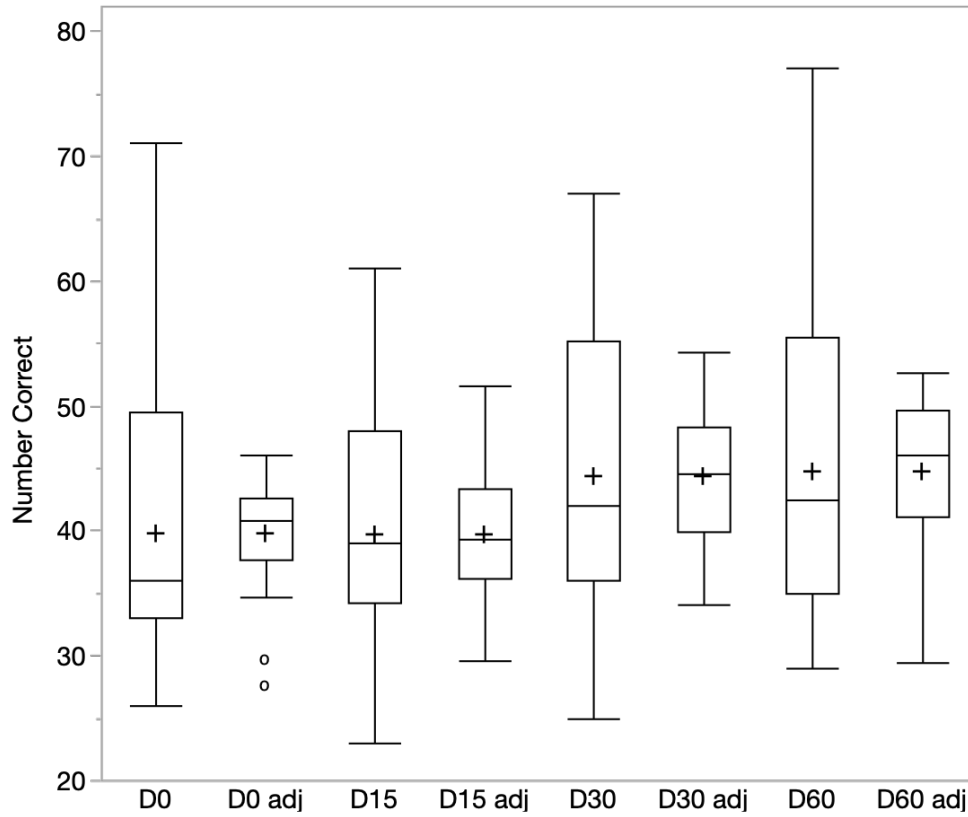


Figure 7 ■ Box plots of the raw score and adjusted scores (adj) as a function of dose.



lows: for each set of coefficients, the first score is multiplied by the first coefficient, the second score by the second, etc. Then these six products were summed. The data for one of the subjects was 12.59, 10.45, 7.68, 7.30, 12.96, and 9.70. The sum of cross products of these scores with the coefficients for the linear trend results in a score on the linear component of -0.9 . See Maxwell and Delaney (2004) for further details.

Figure 10 plots these two variables as a function of group. This figure shows a sizable group difference in linear trends with the median of the experimental group being only slightly below the 75th percentile of the control group and the 25th percentile of the experimental group being only slightly below the median of the control group. The size of the group difference is similar for the quadratic trend except that the quadratic trend is larger for the control group. No outside values or skew are apparent in this figure.

Summary and Conclusions

There are a variety of reasons descriptive distributional information is important. First, it allows the assessment of any violations of the assumptions made in the inferential statistics or other aspects of distributions that may call into question the validity of the conclusions. It is important to display descriptive distributional information even when such problems are not apparent so that readers can make their own assessments. Second, just as Stephen Gould famously wrote “The median isn’t the message” (Gould, 2013) a strong case can be made that the mean is not the entire message either. Shapes of distributions can be theoretically important in their own right and not just relevant to assumptions required to test for mean differences. For example, a finding that there is a bimodal distribution could be theoretically interesting. Finally a graph of a distribution may reveal that an effect may be different for different portions of the population. One example can be found in the data portrayed in Figure 5 which suggests the critical interaction effect is greater for subjects who, overall,

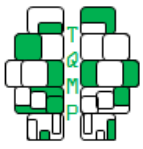
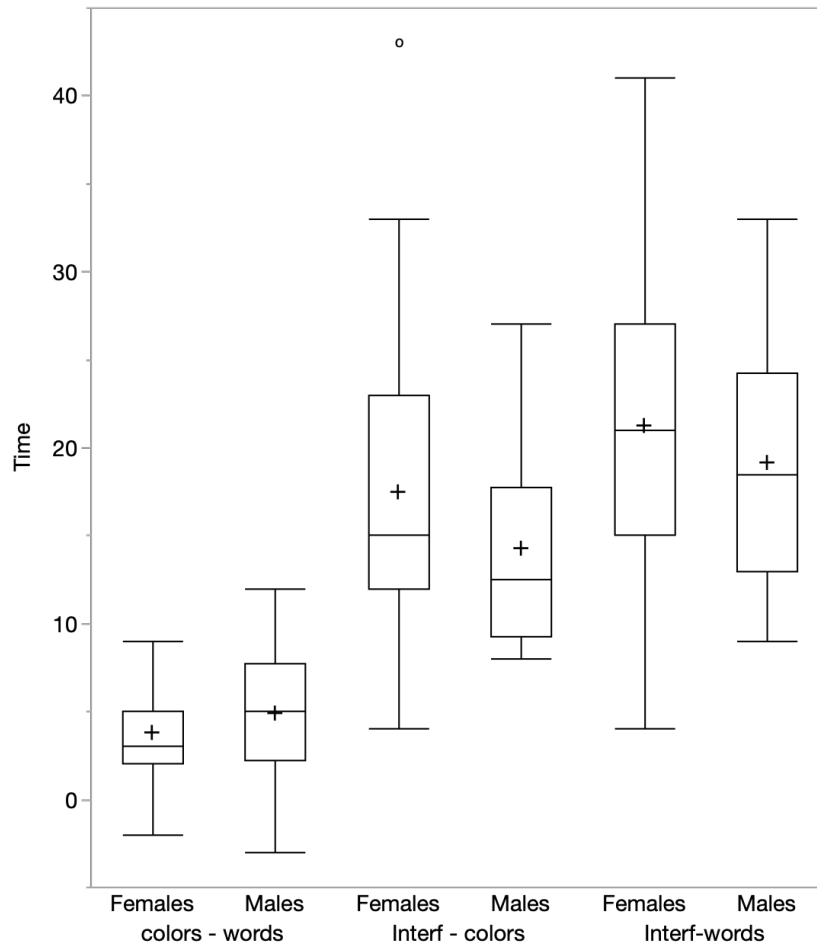


Figure 8 ■ Box plots showing difference scores as a function of gender.



respond quickly and those who respond slowly than those whose overall response times are closer to the mean.

The main thesis of this article is that descriptive distributional information in within-subjects designs should be shown in a way that is not obscured by between-subjects variability. The graphs shown here are examples of the types of graphs that serve this purpose but clearly are not the only graph types that do so.

The choice of graph type depends on many factors. Comparative graphs can be very effective when there are relatively few observations. However, the number of observations that can be displayed effectively depends on the nature of the data. If the differences between conditions are very similar across subjects so that the lines are approximately parallel, then a graph with a large number of subjects can be apprehended easily. However, if the differences are not consistent, with some being positive and

some being negative, the graph will have numerous intersecting lines and will not communicate the results well.

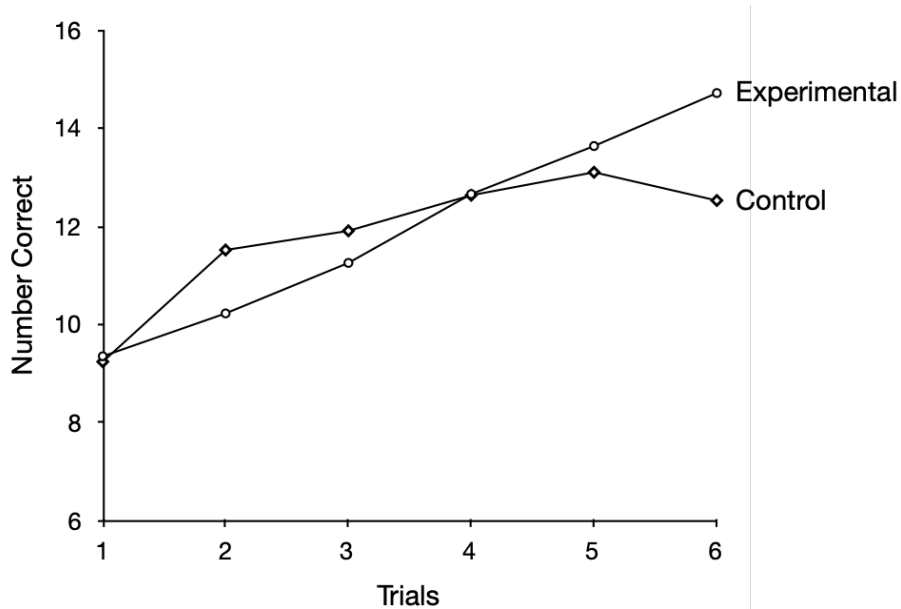
Plotting difference scores is a good alternative to comparative graphs when there are many observations. Examples given here include jittered dot plots and box plots. Other possibilities include stem and leaf displays, histograms, and density plots. The latter two of these are usually the better choices with very large datasets.

Sum-difference graphs portray more information than plots of difference scores and should routinely be created. If a sum-difference plot does not reveal anything of theoretical importance not found in graphs of difference scores, the plot of difference scores would typically suffice in a published report.

The number of pairwise comparisons and associated graphs is large when there are many levels of a within-subject variable. However, if the variances of pairwise dif-



Figure 9 ■ Line graph of the means as a function of condition.



ference are similar, the number of graphs necessary to portray the distributions can be markedly reduced by adjusting the scores in the manner shown previously. When the levels of the within-subjects variable are numerically ordered, graphing of trend components rather than all pairwise comparisons can often communicate the most relevant aspects of the distributions effectively. Complex comparisons such as components of interactions can be portrayed similarly.

The principle that variables controlled in the statistical analysis should also be controlled in the graphs is applicable to designs other than within-subjects designs. For example, in analysis of covariance (ANCOVA), the variance accounted for by the covariate should not be portrayed as random error in graphs as it would be if the raw scores were plotted. Since significance tests in ANCOVA are tests of adjusted means with the error term based on the variance remaining after controlling for the covariate, graphs should show adjusted means and variability remaining after controlling for the covariate. This can be done by plotting adjusted scores computed as follows: The first step in computing adjusted scores is to save the residuals from a model containing both groups and the covariate. These residuals are not suitable for graphing, however, since the effect of groups is controlled making the means of all groups the same. The next step, therefore, is to add the adjusted mean (sometimes called least squares mean or estimated marginal mean) of each group to the residual for

each subject in the group. Adjusted means are routinely computed by major statistics software such as R, SPSS, and SAS. If the covariate explains a substantial amount of variance, then variability shown in the graph of adjusted scores will be considerably lower than in a graph of raw scores.

There are similar issues in regression analysis. Although in simple regression, a scatterplot that includes the regression line is a very informative way to display the relationship between the predictor variable and the criterion, a scatterplot of a single predictor and the criterion has two problems in a multiple regression analysis: (1) the multiple regression equation contains the partial regression slope for the predictor variable under consideration rather than the simple regression slope represented in the line in the scatterplot and (2) variance explained by other predictor variables in the model would be represented as random error in the scatterplot. A partial regression plot (Velleman & Welsch, 1981) is a better way to show the relationship between a predictor and the criterion in multiple regression. The basic idea is to create a plot in which the effects of all predictor variables except the one under consideration are controlled so that the slope of the line in the scatterplot is the partial slope and the error variance is correctly displayed. Specifically, a partial regression plot of predictor $X(i)$ is constructed by (a) finding the residuals in Y after being regressed on all variables except $X(i)$, (b) finding the residuals in $X(i)$ after being regressed on all

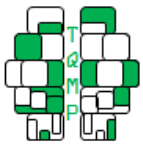
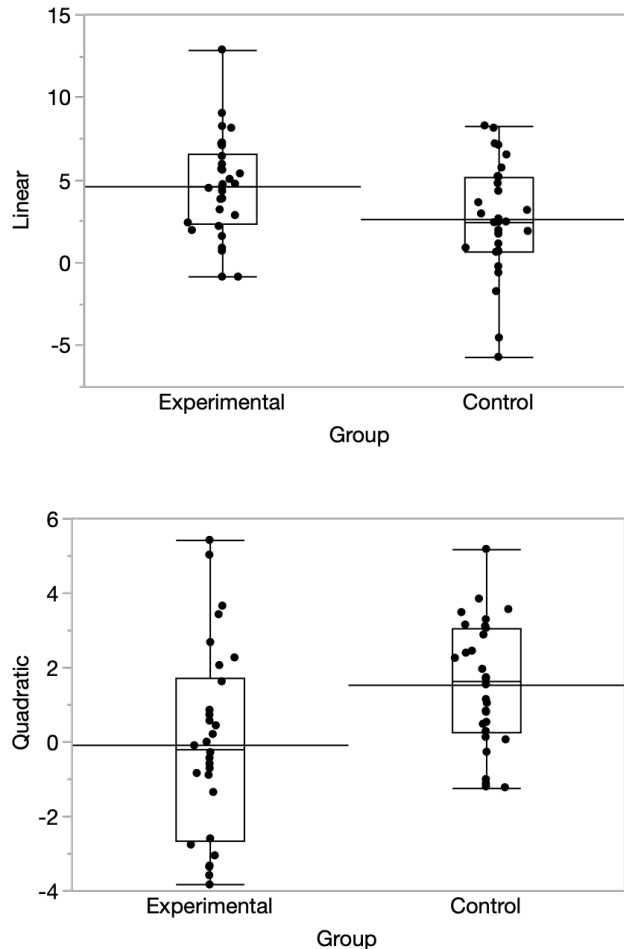


Figure 10 ■ A version of box plots showing linear and quadratic trend components as a function of condition. The horizontal lines going through the boxes represent means whereas the lines within the boxes represent medians. The points are jittered to reduce overlap.



predictor variables except $X(i)$, (c) adding the mean of Y to its residuals, (d) adding the mean of $X(i)$ to its residuals, and (e) creating a scatterplot of the Y and X adjusted variables that includes a regression line. The slope of this regression line will equal the regression coefficient for $X(i)$ in the multiple regression equation whereas the slope of the simple regression of Y on $X(i)$ using the raw scores would not.

Showing more data in graphs does not mean that you invariably learn something you would not know from the means themselves. However, even if nothing important is revealed, the very fact that the readers have the opportunity to see for themselves that there are no data irregularities is important. This includes whether assumptions are severely violated and whether there are very influential

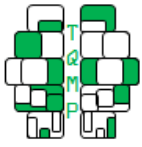
data points.

Tufte's (2001) first principle of graphical excellence is to show the data. As noted by Wilkinson and the APA Task Force on Statistical Inference (1999), failure to show data often hinders scientific evaluation. The graph types illustrated here show data from within-subjects designs in a way that controls for between-subjects variation. Moreover, they are relatively easy to produce, and, in many cases, take approximately the same amount of space as the much more common but information-poor bar charts of means.



References

- Altman, D. G., & Bland, J. M. (1983). Measurement in medicine: The analysis of method comparison studies. *The Statistician*, 32, 307–317. doi:[dx.doi.org/10.2307/2987937](https://doi.org/10.2307/2987937)
- Anderson, C. A., Benjamin, A. J., & Bartholow, B. D. (1998). Does the gun pull the trigger? automatic priming effects of weapon pictures and weapon names. *Psychological Science*, 9, 308–314.
- Baguley, T. (2012). Calculating and graphing within-subject confidence intervals for anova. *Behavior Research Methods*, 44, 158–175. doi:[10.3758/s13428-011-0123-7](https://doi.org/10.3758/s13428-011-0123-7)
- Bakeman, R., & Mearthur, D. (1996). Picturing repeated measures: Comments on loftus, morrison, and others. *Behavior Research Methods, Instruments, & Computers*, 28, 584–589. doi:[10.3758/BF03200546](https://doi.org/10.3758/BF03200546)
- Cleveland, W. S. (1994). *The elements of graphing data (2nd edition)*. Summit, NJ, USA: Hobart Press.
- Correll, M., & Gleicher, M. (2014). Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Transactions on Visualization and Computer Graphics*, 20(12), 2142–2151. doi:[10.1109/TVCG.2014.2346298](https://doi.org/10.1109/TVCG.2014.2346298)
- Cousineau, D., & O'Brien, F. (2014). Error bars in within-subject designs: A comment on baguley (2012). *Behavior Research Methods*, 46, 1149–1151. doi:[10.3758/s13428-013-0441-z](https://doi.org/10.3758/s13428-013-0441-z)
- Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals, and how to read pictures of data. *American Psychologist*, 60, 170–180. doi:[dx.doi.org/10.1037/0003-066X.60.2.170](https://doi.org/10.1037/0003-066X.60.2.170)
- Duke, S., Bancken, F., Crowe, B., Soukup, M., Botsis, T., & Forshee, R. (2015). Seeing is believing: Good graphic design principles for medical research. *Statistics in Medicine*, 22, 3040–3059. doi:[10.1002/sim.6549](https://doi.org/10.1002/sim.6549)
- Gould, S. J. (2013). The median isn't the message. *Virtual Mentor*, 15, 77–81. doi:[10.1001/virtualmentor.2013.15.1.mnar1-1301](https://doi.org/10.1001/virtualmentor.2013.15.1.mnar1-1301)
- Harris, R. (1999). *Information graphics: A comprehensive illustrated reference*. Atlanta, GA: Management Graphics.
- Krzywinski, M., & Altman, N. (2014). Visualizing samples with box plots. *Nature Methods*, 11, 119–120. doi:<https://doi.org/10.1038/nmeth.2813>
- Lane, D. M., & Sandor, A. (2009). Designing better graphs by including distributional information and integrating words, numbers, and images. *Psychological Methods*, 14, 239–257. doi:[10.1037/a0016620](https://doi.org/10.1037/a0016620)
- Larsen-Hall, J. (2017). Moving beyond the bar plot and the line graph to create informative and attractive graphics. *Modern Language Journal*, 101, 244–270. doi:[doi:10.1111/modl.12386](https://doi.org/10.1111/modl.12386)
- Loftus, G. R. (1995). Data analysis as insight: Reply to morrison and weaver. *Behavior Research Methods, Instruments, & Computers*, 27, 57–59. doi:doi.org/10.3758/BF03203621
- Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin and Review*, 1, 476–490. doi:[10.3758/BF03210951](https://doi.org/10.3758/BF03210951)
- Marmolejo-Ramos, F., & Matsunaga, M. (2009). Getting the most from your curves: Exploring and reporting data using informative graphical techniques. *Tutorials in Quantitative Methods for Psychology*, 5, 40–50. doi:[10.20982/tamp.05.2.p040](https://doi.org/10.20982/tamp.05.2.p040)
- Martinez, E. (2015). Description of continuous data using bar graphs: A misleading approach. *Revista da Sociedade Brasileira de Medicina Tropical*, 48, 1–11. doi:[10.1590/0037-8682-0013-2015](https://doi.org/10.1590/0037-8682-0013-2015)
- Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective (2nd ed.)*. Lawrence erlbaum associates publishers. New Jersey: Mahwah.
- McNeil, D. (1992). On graphing paired data. *The American Statistician*, 46(4), 307–311. doi:[10.1080/00031305.1992.10475915](https://doi.org/10.1080/00031305.1992.10475915)
- Millard, S. P. (2013). *Envstats: An r package for environmental statistics*. New York: Springer. Retrieved from <http://www.springer.com>
- Pearson, D. A., Santos, C. W., Casal, C. D., Lane, D. M., Jerger, S. W., Roache, J. D., ... Cleveland, L. A. (2004). Treatment effects of methylphenidate on cognitive functioning in children with mental retardation and adhd. *Journal of the American Academy of Child and Adolescent Psychiatry*, 43, 677–685. doi:[10.1097/01.chi.0000124461.81324.13](https://doi.org/10.1097/01.chi.0000124461.81324.13)
- Schriger, D. L. (2017). Graphic paired data: A tutorial. *Annals of Emergency Medicine*, 71, 239–246. doi:[10.1016/j.annemergmed.2017.05.033](https://doi.org/10.1016/j.annemergmed.2017.05.033)
- Tufte, E. R. (2001). *The visual display of quantitative information (2nd ed.)*. Cheshire, CT: Graphics Press.
- Velleman, P. F., & Welsch, R. E. (1981). Efficient computing of regression diagnostics. *The American Statistician*, 35, 234–242. doi:[10.1080/00031305.1981.10479362](https://doi.org/10.1080/00031305.1981.10479362)
- Wallenstein, S., & Fleiss, J. L. (1979). Repeated measurements analysis of variance when the correlations have a certain pattern. *Psychometrika*, 44, 229–233. doi:[doi:10.1007/BF02293973](https://doi.org/10.1007/BF02293973)
- Weissgerber, T., Milic, N., Winham, S., & Garovic, V. (2015). Beyond bar and line graphs: Time for a new data presentation paradigm. *PLoS Biology*, 13, 1–10. doi:[10.1371/journal.pbio.1002128](https://doi.org/10.1371/journal.pbio.1002128)



Wilkinson, L., & the APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594–604. doi:<http://dx.doi.org/10.1037/0003-066X.54.8.594>

Wright, T., Klein, M., & Wiecek, J. (2019). A primer on visualizations for comparing populations, including the issue of overlapping confidence intervals. *The American Statistician*, 73, 165–178. doi:[10.1080/00031305.2017.1392359](https://doi.org/10.1080/00031305.2017.1392359)

Appendix: Data Sources


Anderson et al. Weapons Effect. http://onlinestatbook.com/2/case_studies/guns.html. Note that in the original article, the groups were specified by the kind of target word followed by the kind of priming word so that, for example, AW represented an aggressive word primed by a weapon word. To make the coding consistent with the temporal order of events, the code letters are reversed in this article so that WA represents the condition in which a weapon word prime was followed by an aggressive target word.

Pearson et al. ADHD Study. http://onlinestatbook.com/2/case_studies/adhd.html

Stroop Effect. http://onlinestatbook.com/2/case_studies/stroop.html

Repeated Measures with linear and quadratic trends. See Supplemental material on journal website

Open practices

 The *Open Data* badge was earned because the data of the experiment(s) are available on the [journal's web site](#).

Citation

Lane, D. M. (2019). Graphing within-subjects effects. *The Quantitative Methods for Psychology*, 15(3), 174–187. doi:[10.20982/tqmp.15.3.p174](https://doi.org/10.20982/tqmp.15.3.p174)

Copyright © 2019, Lane. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Received: 25/05/2019 ~ Accepted: 04/09/2019