# Memo

To: Colleagues
From: Roger Bakeman
Date: June 20, 2021
Re: Outliers, transformations, nonparametric statistics, and recodes:
When if ever to modify data before analysis

I call them zombie rules; you can't kill them. In articles I am asked to review and in conversations with colleagues and students, I occasionally encounter statements like:

- Our data weren't normally distributed so we transformed them to make them normal.
- Outliers are scores 3 standard deviations or more beyond the mean.
- Some data points were outliers, so we deleted them.
- Our data weren't normally distributed, so we used nonparametric statistics.

Often the source of such statements is a memory of something "taught" in a long-ago statistics class, but as psychologists we all know about the fallibility of human memory.

The principle I would promote is **fiddle seldom**. Certainly, as a preliminary to analysis, examine how data are distributed (data screening), but modify data before analysis only when accompanied with a strong, explicit, convincing rationale and knowledge of the consequences. The problem, as I see it, is the too mechanical application of "rules," in the often naïve belief that doing so will automatically fix the problem. Would that the world were so simple.

**Why the Obsession with Normally Distributed Data?**

A statement like, "I transformed my data to make them normal," is simply false. The usual transformations can make a distribution less skewed, but they cannot make it normal. Yet, even Tabachnick and Fidell in their popular book, *Using Multivariate Statistics* (various editions, 1st 1983, 2nd 1989, 3rd 1996, ... 7th 2019), in an otherwise excellent section in Data Screening, label a figure as "Original distributions and common transformation to produce normality."

A normal distribution is one that closely approximates a specific, not especially simple mathematical function. It involves pi ($\pi$), which equals approximately 3.1416, the base of the natural logarithm (*e*), which equals approximately 2.7183, and a negative exponent. For standardized scores (symbolized *Z*; mean = 0, standard deviation = 1), the function reduces to the *standard normal*:

$$Y = \frac{1}{\sqrt{2\pi}} e^{-\frac{Z^2}{2}}$$

The underlying circumstances that produce normally distributed numbers are not difficult to understand. Whenever the thing measured is the result of *multiple, independent, random* causes, the scores will be normally distributed. This is not mysterious (as some French 19th century mathematicians thought; Stigler, 1986), but simply a straightforward mathematical consequence. Because so many phenomena we study are indeed the result of multiple, independent, random genetic and environmental influences, normally distributed scores are frequently encountered. Or would be, if our samples were large enough, which they rarely are.

Why care if scores we intend to analyze are “normally” distributed? The concern probably reflects something we were taught in introductory statistics and that is now a dim memory. As a typical example, let me quote a classic text, Keppel and Saufley (1980), *Introduction to Design and Analysis: A Student’s Handbook*. When discussing evaluation of the *F*-ratio in the context of an analysis of variance, they write:

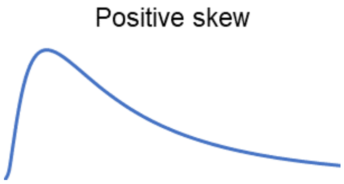
> We start with the notion of treatment populations—extremely large numbers of individuals randomly subjected to the different treatment conditions. There is a different treatment population for each condition ... The null hypothesis states that the population treatment means are equal.” ... [In order to use the theoretical sampling distribution of the *F* ratio to assign a *p* value to our results] we make three assumptions regarding the individual scores of the subjects present in these hypothetical treatment populations:
>
> 1. That they distribute themselves normally
> 2. That they show the same degree of variability from treatment population to treatment population [homogeneity of variance]
> 3. That they are independent from one another both within each treatment population and across populations

Note, first, that these assumptions apply to populations, not to the samples actually selected, and second, matter only if you wish to assign a *p* value to your results. However, as most authors do, Keppel and Saufley continue:

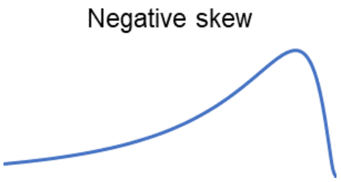
> ... to what extent do deviations from them [these assumptions] jeopardize the procedures we employ in evaluating the null hypothesis? ... [when not met] however, it appears that even relatively severe deviation from the conditions assumed [including random selection of the sample from the treatment populations] have little effect on the evaluation process.

This leads me and others to conclude that even relatively severe deviations of sample data from an ideal normal distribution are not sufficient reason to transform data before analysis.

### How About Outliers and Skew?



When our colleagues and others say that a distribution is not normal, often they mean simply that it is skewed, either **positively** (a preponderance of low numbers piling up on the left with a long tail to the few high numbers on the right) or **negatively** (a preponderance of high numbers piling up on the right with a long tail to the few low numbers on the left). Often the skew is accompanied by a few **extreme scores**, either unusually small scores, or unusually large scores, relative to the other scores in the distribution.



Extreme scores are often called outliers, but outliers are defined in a variety of ways and sometimes it is not easy to say whether a score is an outlier or simply the tail of a skewed distribution. Extreme scores can be “legitimate” data.

### *Outliers*

Tabachnick and Fidell (1983 ... 2019) define outliers as “cases with such extreme values on one variable or a combination of variables that they distort statistics.” Noteworthy is that they provide no simple, mechanical definition or rule. Their discussion of outliers is well worth reading but here is a brief summary: Data should be changed if caused by a typo. A case should be deleted if it didn’t meet inclusion criteria. An extreme score might be modified (made less extreme) if it distorts results. Otherwise, fiddle seldom; but more later.

### *Portraying Distributions*

The first step in approaching any data set, a necessary step before any analysis is even contemplated, is **data screening.** (Again, I recommend Tabachnick and Fidell; their discussion of this topic is also well worth reading). Pictures help. Best by far is a bubble plot but a close second is a box-and-whisker plot. SPSS’s Chart Builder makes plots acceptable for inspection if not quite of publication quality. Excel makes acceptable box-and-whisker plots. But all let you see the shape of the distribution, identify extreme scores, and note any other quirks.

Tukey’s (1976) definition of an **extreme score**—a score that exceeds 1.5 times the interquartile range (IQR) is particularly useful

It is based on percentiles, and so is unaffected by how extreme sores are.

(Note, 1.5 is the standard multiplier, but other values have been suggested; e.g., Hoaglin & Iglewicz, 1987).

In Fig. 1, extreme sores are portrayed with circles. The whiskers are the lowest and highest scores that are not extreme (41 and 55) although 1.5 × IQR would define scores 40.25–62.125 as not extreme.

The solid horizontal line in Fig. 2 is the mean, which shows how the mean can be influenced by extreme scores in a skewed distribution.

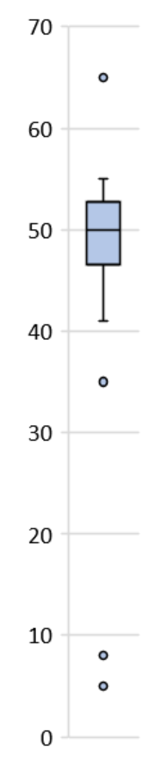


*Figure 1.* Box-and-whisker plot; made with Excel.

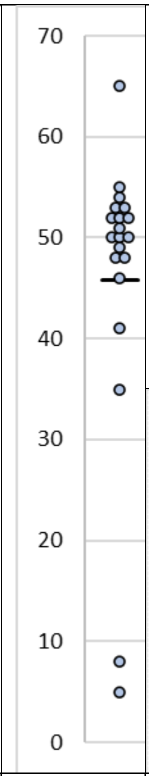
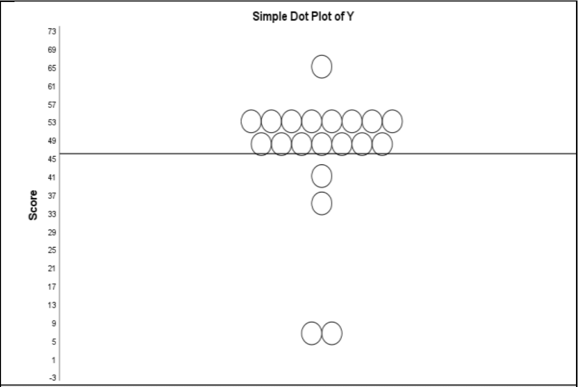


*Figure 2.* Bubble plot; hand-built with Excel.



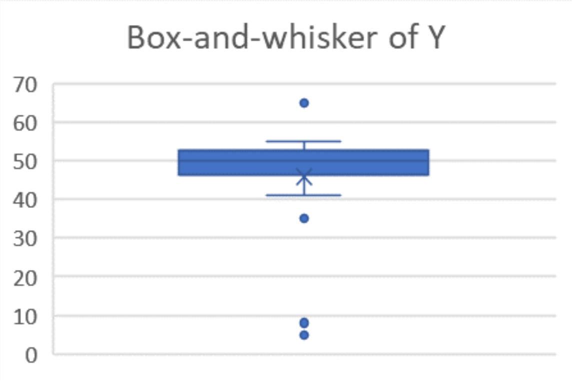


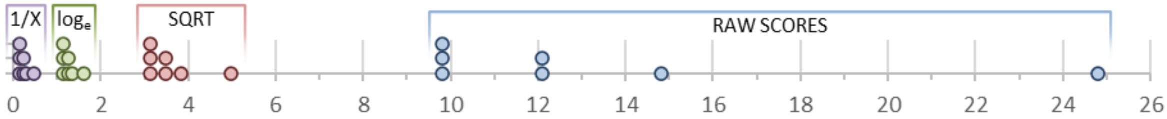
*Figure 3* (top). SPSS bubble plot. The Increment was specified as 1, but defaulted to 4, thus data points were lumped. Excel gives more control.
*Figure 4* (bottom). SPSS box-and-whisker plot; with copy-and-paste to Excel, it can be formated to look like Figure 1.

### *Transformations to Reduce Skew*

In discussing *Common Data Transformations*, Tabachnick and Fidell begin by noting that "Although data transformations are recommended as a remedy for outliers and failures of normality, linearity, and homoscedasticity, they are not universally recommended," but then go on write:

> Our recommendation, then, is to consider transformation of variables in all situations unless there is some reason not to. ... With almost every data set where we have used transformation, the results of analysis have been substantially improved. This is particularly true when some variables are skewed and others are not, or variables are skewed very differently prior to transformation. However, if all the variables are skewed to about the same moderate extent, improvements of analysis with transformation are often marginal."

Perhaps it is these statements that have encouraged our colleagues to transform data, but I have always found them a bit puzzling. Rarely have I seen "substantial improvement" in data sets I have analyzed after transforming variables. In any event, here is a simple demonstration showing how the common transformations Tabachnick and Fidell mention work.



From right to left, are (a) a set of seven positively skewed scores, (b) the **square root** of the raw scores, (c) the **natural log** of the raw scores, and (d) the **reciprocal** of the raw scores (i.e., 1/x). Note how each successively stronger transformation compresses the scores more—but the basic shape of the distribution is unchanged. Again from right to left, the standardized skews of these data sets are 2.60, 2.35, 2.06, and 1.83. A principle of transformation is that the weakest one that meets criterion should be used. If a standardized skew that exceeded 2.58 ($p < .01$, two-tailed, normal distribution) were our criterion, then the square root transformation would suffice. If one that exceeded 1.96 ($p < .05$, two-tailed, normal distribution) were our criterion, then the reciprocal would be used. (Some prefer a less stringent criterion, e.g. 3.29, $p < .001$.) These transformations reduce skew, as shown. If used, the criterion should be stated along with a statement that the transformation selected was the weakest to meet that criterion. A statement like, "our data were skewed so we used an arcsine transformation," is not sufficient.

If tempted by transformation, first run your analyses on untransformed then on transformed data. Do the results differ in any meaningful way? If not, report results for the untransformed data and note the lack of difference with transformed data. If yes, report results for the transformed data but state how they differed from the untransformed results.

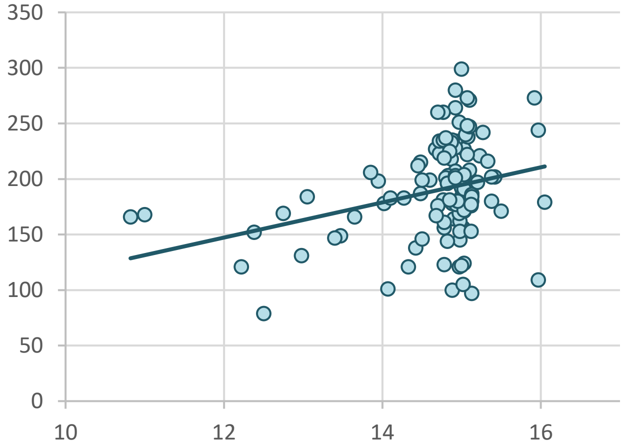
However, **for basic descriptive statistics** (means, medians, standard deviations, box-and-whisker plots, etc.) **always report values based on untransformed data**. Readers who want to compare your results to theirs or to other results in the literature should have raw score statistics readily available without needing to reverse transform. Report medians if they are a better indicator of central tendency than means. Portray variability with box-and-whisker plots; standard deviations with their assumption of unimodal symmetry are not that helpful.

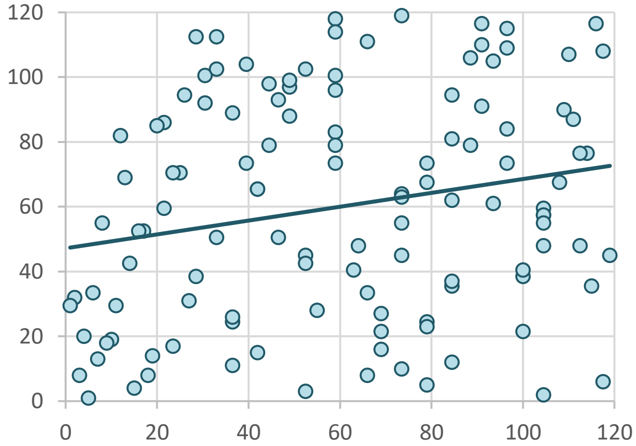
**Nonparametric Statistics: A Rank-Order Transformation**

The common transformations described by Tabachnick and Fidell belong to the world of parametric statistics with its assumption that scores are measured on a continuous, equal-interval scale (i.e. real numbers with fractional parts). And even if sores are initially measured on an integer, equal-interval scale, these transformations produce real numbers.

As you can see from the previous figure, these transformations reduce the distances between scores, reducing most for the highest score, successively less for lower scores (for positively skewed distributions; the reverse for negatively skewed ones). It's as though the number line was first written on a stretched rubber sheet and then the stretch was relaxed, more so for each successively stronger transformation.

Nonparametric statistics are often offered as another solution for badly distributed scores. But nonparametric statistics are just another transformation. Consider the Spearman rank-order correlation. To compute it, first replace the raw scores with their ranks, then apply the standard Pearson product-moment formula. Like the transformations just described, nonparametric statistics alter the distance between score values; unlike them, they make all distances equal. Here is an example:





The figure on the left shows a scatterplot for 119 scores (*r* = .30, *p* < .001) adapted from an existing data set. The figure on the right shows a scatterplot for their ranks (ρ = .21, *p* = .020). In this case, the correlation based on raw scores is barely moderate, although larger than the weak correlation based on ranks. However, the two are not all that dissimilar, which is typically the case when Pearson and Spearman correlations are compared. The question for the researcher is, which scatterplot is more informative. Does it make sense to regard the distance between data points on the left and right as equivalent to the distance between the data points clustered so tightly in the middle?

**Summary.** If you choose to use the standard nonparametric statistics—the Spearman rho for the association between two variables, the Mann-Whitney U test for the difference between medians of independent groups, or the Wilcoxon signed-rank test for the difference between medians of related groups—be aware that these statistics transform raw scores into their ranks. Does it make sense to regard all differences between adjacent scores as equivalent?

## Recodes: A Transformation Worth Considering

Generally I ague that altering data before analysis should be done seldom and that, when done, requires strong justification accompanied by a comparison with the unaltered data results. It should never be done on the assumption that doing so fixes some problem without evidence that it actually does so (mea culpa; we evolve). However, there are two exceptions: binary and ordinal recodes.

### *Binary Recodes*

A distribution may be skewed because a sizable proportion—30% or more, say—are all 0, if 0 is the lowest score possible; or are all 100, if 100 is the highest score possible. Conceptually, such distributions might better be treated as binary. For example, 0 could indicate none and 1 some of whatever is measured; or 1 could indicate the highest score achievable, with 0 indicating that a score fell short. In such cases, the mean can be a misleading summary measure. The percentage of scores coded 0 or 1 is more informative and a binary recode may well be the better score to analyze and interpret. Transformations, including replacing scores with their ranks, won't work well because even after transformation the same sizable proportion of scores will still have the same value.

### *Ordinal Recodes*

A distribution may also be skewed when intervals between successive value are not conceptually equal. Two examples are number of symptoms and years of education. The "distance" between 0 symptoms and 1 is not the same as the distance between 3 and 4 symptoms, which is not the same as the distance between 18 and 19 symptoms. In this case, an ordinal recode that attempts to equate distances is often the best solution. For example, 0, 1, 2–3, 4–6, 7–10, 10–20, over 20 symptoms might be recoded with the ordinal steps as 0–7. Similarly, the distance between 6 and 7 years of schooling is not the same as the distance between 11 and 12. Depending on characteristics of the sample, a recode like 1 = some high school, 2 = finished high school, 3 = some college, 4 = finished college, 5 = some post college, 6 = graduate degree might capture educational differences better than years of education.

The standard parametric statistics assume equal-interval measurement. Sometimes the numbers collected—count variables like number of symptoms is a prime example—only appear to be arranged on an equal-interval scale. It is the task of ordinal recodes to make intervals that are more conceptually equal.

## Recommended Rules

1. Before any analysis, always examine distributions. What is the appropriate level of measurement: binary, ordinal, equal-interval integer, equal-interval continuous?
2. Is a binary or ordinal recode suggested?
3. If any data are altered before analysis, provide a rationale and compare analytic results using unaltered data with results using altered data.
4. Base descriptive statistics on unaltered data, using statistics that take distribution quirks into account. Box-and-whisker plots reveal distributions better than a simple *M* and *SD*.

## Appendix A
## All Correlations Coefficients are the Same

We give different names to correlations, depending on the data types for the two variables:

1. Pearson product-moment, the correlation between two continuous variables.
2. Spearman, the correlation between the ranked scores of two continuous variables.
3. Point biserial, the correlation between a continuous and a binary variable.
4. phi coefficient, the correlation between two binary variables.

Textbooks give different computational formulas for these "different" correlations, and they have different names, which gives rise to the mistaken belief that they are somehow different. In fact, all four types of correlations have a single conceptual formula: The correlation coefficient—any correlation coefficient—is the average cross-product of the standardized scores. Alternatively, it is the ratio of the covariance to the square root of the product of the variances for the two variables:

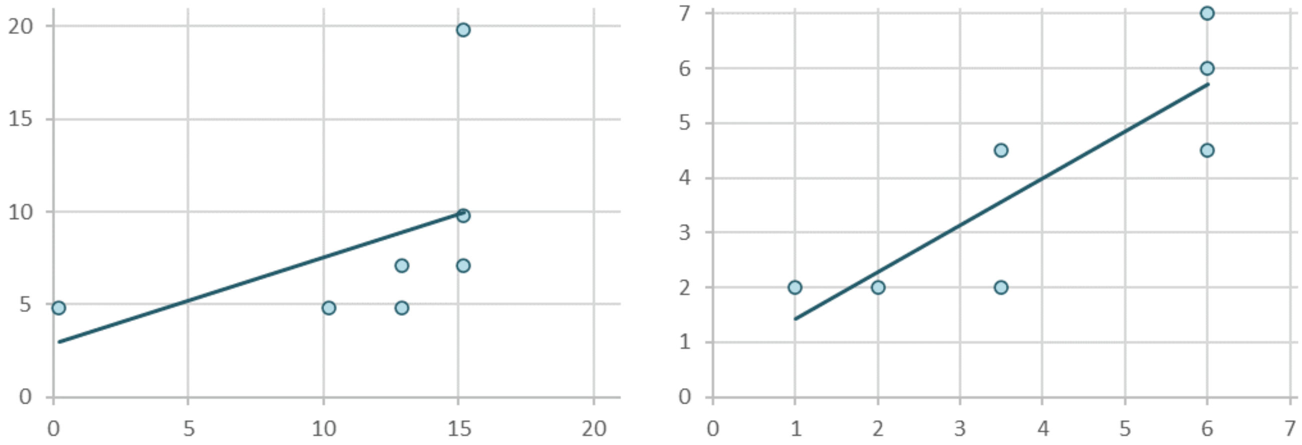
$$r = \frac{\sum(Zx_i Zy_i)}{N} \text{ (when } i = 1 \text{ to } N) \; = \frac{\mathrm{COV}_{xy}}{\sqrt{\mathrm{VAR}_x \mathrm{VAR}_y}},$$

Applying these conceptual formulas to data, no matter the data type, produces the correlation coefficient. We then give the coefficient a different name, depending on the data types involved. Thus, if you replace raw scores with their ranks and ask SPSS to compute a Pearson correlation, you get the same result as if you had asked SPSS to compute a Spearman correlation on the raw data.

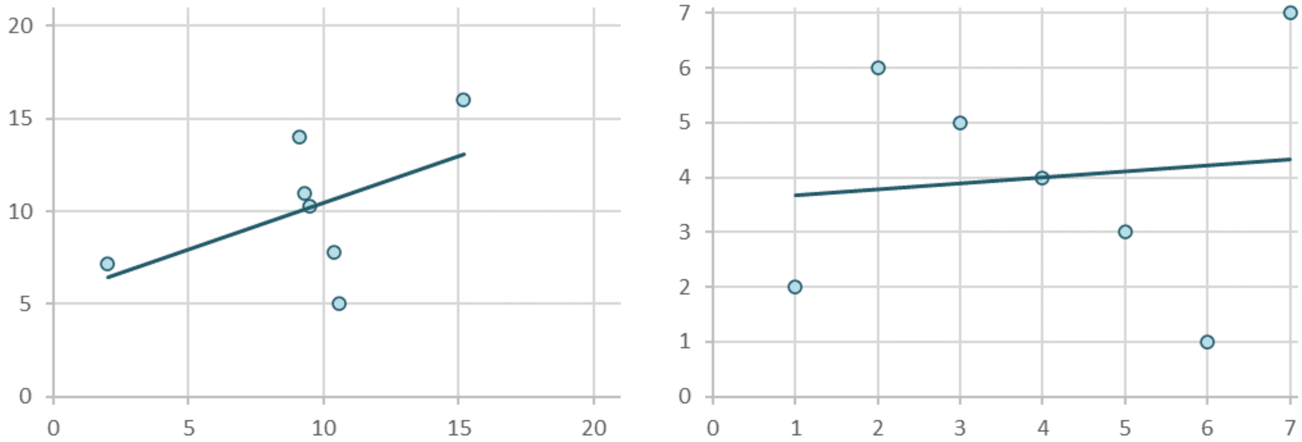
*Note*. A fifth name is the biserial correlation. When the binary variable results from dichotomizing an underlying continuity (e.g., low scores and high scores) the correlation is called biserial. When the binary variable is truly binary (e.g., male and female) the correlation is called point biserial.

## Appendix B
## Correlations for Untransformed and Transformed Data Compared

Sometime small, concocted data sets can illustrate points.



When two variables are skewed in opposite directions, how is their correlation affected by transformation? Here, the X-axis scores are skewed negatively and Y-axis scores are skewed positively (*N* = 7). The left-hand scattergram shows the raw scores, the right-hand one shows their ranks. The three highest scores were the same so their tied rank is 6; the next two highest scores were the same so their tied rank is 3.5. *r* = .47, .44, .41, .41, and .85 for raw scores and for square-root, natural logarithm, reciprocal, and rank-order transformations, respectively. For this example, the standard transformations had little effect on the value of the correlation but the rank-order one (Spearman) did.



When many X-axis scores cluster tightly in the center, but a low X-axis score is associated with a low Y-axis score, likewise for a high score, how is their correlation affected by transformation? Again, the left-hand scattergram shows the raw scores, the right-hand one shows their ranks. *r* = .51, .43, .36, .25, and .11 for raw scores and for square-root, natural logarithm, reciprocal, and rank-order transformations, respectively. For this example, transformation steadily decreased the value of the correlation, most dramatically for the rank-order one.

These examples are extreme and fiddled to make a point, but make me question how useful Spearman correlations are. If considering Spearman correlations, always examine scattergrams.

## References

David C. Hoaglin & Boris Iglewicz. (1987). Fine-tuning some resistant rules for outlier labeling. *Journal of the American Statistical Association, 82*(400), 1147–1149 https://doi.org/10.1080/01621459.1987.10478551

Keppel, G., & Saufley, W. H., Jr. (1980). *Introduction to design and analysis: A student's handbook.* New York: W. H. Freeman.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900.* Cambridge, MA: Belnap Press of Harvard University Press.

Tabachnick, B. G., and Fidell, L. S. (1st ed., 1983; 2nd ed., 1989; 3rd ed., 1996; ... 7th ed., 2019). *Using multivariate statistics*. New York: HarperCollins Publishers. Each edition more costly and heavier than the previous but substantive content little changed, mainly additional material added; e.g., the 3rd edition added a chapter on structural equation modeling.

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