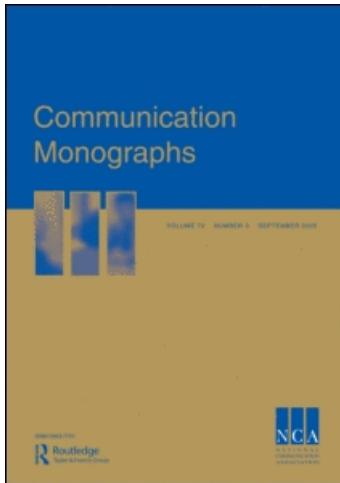


This article was downloaded by: [Cleveland State Univ Libraries]  
On: 28 January 2011  
Access details: Access Details: [subscription number 906603331]  
Publisher Routledge  
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



## Communication Monographs

Publication details, including instructions for authors and subscription information:

<http://www.informaworld.com/smpp/title~content=t713695619>

### The FAQs on Data Transformation

Edward L. Fink

Online publication date: 01 December 2009

**To cite this Article** Fink, Edward L.(2009) 'The FAQs on Data Transformation', *Communication Monographs*, 76: 4, 379 — 397

**To link to this Article:** DOI: 10.1080/03637750903310352

URL: <http://dx.doi.org/10.1080/03637750903310352>

## PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

# The FAQs on Data Transformation

Edward L. Fink

Although we often hear that data speak for themselves, their voices can be soft and sly. (Mosteller, Fienberg, & Rourke, 1983, p. 234)

The job of the data analyst, strangely enough, is to find random error (Winer, 1968). When all systematic variability has been removed from data, the leftover—or residual or disturbance or error—will be random, without pattern; the analyst knows that the analysis is complete when random error has been found.<sup>1</sup> To find random error, data often need to be transformed nonlinearly, and this issue is the topic of this paper.

Data transformation has a long history (e.g., Box & Cox, 1964) and was already in review essays in the 1960s (Kruskal, 1968). Tukey's (1977) volume on exploratory data analysis, with its emphasis on data transformation, created new interests in these techniques; Cohen (1990) stated that “John Tukey's (1977) *Exploratory Data Analysis* is an inspiring account of how to effect graphic and numeric analyses of the data at hand so as to understand *them*” (p. 1310, emphasis in original). Yet to some, exploratory data analysis (EDA), and the data transformations that are an integral part of EDA, appear illegitimate or novel or exotic.

Having taught this topic many times, I am aware of the questions that typically arise concerning data transformation. Thus, I have organized this paper around these questions. So, first:

## Why Do We Transform Data?

### *The Fundamental Link between Measurement and Functional Form*

One of the goals of scientific work is to establish relationships between variables. Consider the simplest case: one independent variable, X, and one dependent variable, Y. Assume that (1) the metrics for these two variables are conventional (i.e., the measurement rules are known and they are considered the standard for use by the

---

Edward L. Fink is a professor in the Department of Communication at the University of Maryland. The author acknowledges the initial work on this topic coauthored with Professor C. L. Bauer of Marquette University and the kind insistence by Professor D. J. Hamble that this topic be included in *Communication Monograph's Issues Forum*. Special thanks go to Professor D. A. Cai for her accurate challenges to what I thought was clear writing. In addition, I wish to thank Professor M. R. Allen and Ms. I. A. Cionea for their many helpful comments, and Professor A. F. Hayes for his thoughtful and thorough reading and suggestions. Unfortunately, any remaining errors are the author's. Correspondence to: Edward L. Fink, Department of Communication, 2130 Skinner Building, University of Maryland, College Park, MD 20742-7635, USA. E-mail: elf@umd.edu

relevant community of scientists), (2) the data are from a relatively large sample, (3) the variables are measured with high reliability, and (4) there is a wide range of values of the independent variable. In that case a scattergram should reveal the functional form of relationship between these two variables, which, if it were hypothesized, could be tested statistically. However, if the first assumption were not made, so that measurement decisions do not reflect convention, the scientist is faced with a daunting task: If the measurement rules for the variables are conventional, the functional form that best represents the relationship between the variables (i.e., best by some prespecified criterion) can be discovered; if the functional form is specified (presumably by strong theory), the metric to express each variable can be discovered, by transforming one or both of the original variables if necessary. However, if neither the functional form nor the measurement rules are specified, the analyst must simultaneously impose both. In the communication discipline the measurement rules are not conventional (see, e.g., Torgerson, 1958, on the distinction among fundamental measurement, derived measurement, and measurement by fiat) and hypotheses are not derived from theoretical principles that require particular functional forms; as a result, to operationalize a theory the communication scholar has to engage in successive attempts at harnessing measurement rules and functional form together. As will be shown below, transforming data encompasses and integrates the establishment of measurement rules and the discovery of the functional form between variables.

### *Transforming for Determining Typical Values*

When we analyze data we are interested in two things: description and inference. The most basic description of data involves three features: *location*—where a variable is centered (also called the typical value or the measure of centrality), *spread*—the variability of a variable, and *association*—the relationship between one variable or set of variables and another variable or set of variables. Conventional measures of location include the mean, the median, and the mode; conventional measures of spread include the interquartile range, the standard deviation, and the variance; and conventional measures of association include Pearson's product-moment correlation coefficient (and derivatives, such as the partial correlation, the part correlation, and the multiple correlation), Spearman's rho, and regression coefficients (partial or simple, standardized or unstandardized).

Location—the central tendency in the data—is the most fundamental thing to observe about a variable. Although formulas differ, conceptually the first step in computing a correlation, a standard deviation, a variance, a regression coefficient, or a factor loading is the subtraction of the mean, which is assumed to be the typical value. And we typically assume that the location of a variable is readily summarized by the mean. Furthermore, we assume, usually implicitly, that our data are well behaved, which results in the variable's distribution having the mean, the median, and the mode at about the same place in the variable's frequency distribution.<sup>2</sup> Now, let's examine when that is the case.

Consider Table 1. Assume that the  $x$ -axis for all the graphs in this table is the value of a variable, and the  $y$ -axis is its corresponding frequency. Although the distribution in Cell 1 need not be normal,<sup>3</sup> the mean, median, and mode are all approximately at the same place. Our notion of the mean as the variable's location is reasonable in this case: As we move away from the mean, positively or negatively, the data become less frequent, as if suggesting that the mean was the "target" and that the departures from the mean are unsystematic errors or random disturbances. The mean is a reasonable summary statistic or parameter for these data.

In Cell 2, the mean, median, and mode are at different locations, and it is unclear whether any single measure of central tendency coincides with our notion of "typical value." In Cell 2 each graph represents a skewed variable: The variable is negatively skewed in 2a and is positively skewed in 2b. If we wish to summarize a single variable (a "single batch" of data in the parlance of data explorers) by its typical value, it is suggested that we transform the variable. Thus, one goal of data transformation is to make the different measures of typical value coincide. The simplest type of transformation with unimodal and skewed data is the *single-bend family of transformations*, as follows:

**Table 1** Frequency Distributions of Types of Variables

Modality	Symmetry	
	Symmetric	Asymmetric
Unimodal	(1) 	(2a) 
Multimodal	(3) 	(4) 

$$Y^* = (Y + k)^{(\lambda)}, \text{ such that if } \lambda = 0, Y^* = \ln(Y + k), \text{ and if } \lambda \neq 0, \\ Y^* = (Y + k)^\lambda, \quad (1)$$

where  $Y$  is the original variable,  $Y^*$  is the transformed variable,  $\ln$  is the natural logarithm, and  $k$  is a constant.<sup>4, 5</sup> This transformation is called a single-bend transformation because when  $\lambda \neq 1$  (i.e., when a nonlinear transformation is employed), the graph of  $Y^*$  on  $Y$  is monotonic with a single bend (also referred to as a *one-bend* transformation in Cohen, Cohen, West, & Aiken, 2003).

The single-bend transformation, also called a power transformation, is a version of the Box-Cox transformation (Box & Cox, 1964; Fox, 1997; Hutcheson & Sofroniou, 2006; Montgomery, 2009; Whistler et al., 2004). If data are positively skewed, a  $\lambda < 1$  will typically make the distribution more symmetric, and if the data are negatively skewed, a  $\lambda > 1$  will typically make the distribution more symmetric.

Recall that we are interested in the location of the variable; employing this transformation creates a new variable ( $Y^*$ ) that is likely to be relatively symmetric, the mean (or median or mode) becomes a reasonable typical value.

### *Transforming to Create a Useful Metric*

Transformation of a variable reflects something else about it. Let's consider some positively skewed variables. (The corresponding argument may be made for negatively skewed variables.) Examples are personal income, the size of freely formed social groups, the size of land masses, the number of messages people send (or receive) each day, and the level of intimacy one has with the members of one's social network (assuming our scale was not bound at the upper end, to be discussed below). Notice that, for all these variables, differences in the variable at the low end are greater, in a philosophical and practical sense, than the same differences at the upper end. For example, the change in group processes that occurs when a group goes from three to four members is greater than when a group goes from ten to eleven members, although both are one-unit differences. Similarly, the difference in life style brought about by one's income going from \$20,000 per year to \$40,000 per year is much greater than the difference if one's income went from \$120,000 per year to \$140,000 per year, although both differences are \$20,000. It is generally true that, for positively skewed variables, differences at the low end have a greater impact than equal unit differences at the upper end. When we employ a single-bend transformation, in addition to creating symmetry (thereby making the variable's location meaningful), it is likely that we make the unit differences for the transformed variable reflect the variable's causes and effects in a more reasonable way. In other words, transforming to create a reasonable metric helps create a useful functional form of relationship among or between the variables.

### *Transforming as Differential Stretching and Shrinking*

To show how the single-bend transformation works, consider Table 2. The hypothetical data are modestly positively skewed (skewness = 1.87), but the mean

**Table 2** Hypothetical Data for a Positively Skewed Variable and the Effect of a Transformation

Data	Natural logarithm of the data
0.5	−0.69
1	0.00
2	0.69
3	1.39
10	2.30
20	3.00
50	3.91
100	4.61
Mean (standard deviation)	23.44 (35.11)
Median (interquartile range)	7.00 (41.25)
Median/mean	30%
Skewness/(standard error of skewness) = standardized skewness	1.87/(0.75) = 2.49
Kurtosis/(standard error of kurtosis) = standardized kurtosis	3.18/(1.48) = 2.14
	0.09/(0.75) = 0.12
	−1.29/(1.48) = −0.87

Computations are to four decimals, but results are rounded to two decimal places.

and the median are very different: The median is only 30% of the mean, and it is 47% of a standard deviation from the mean. However, when the variable is transformed by using the natural logarithm (i.e., in reference to Equation 1,  $\lambda=0$  and  $k=0$ ), the skewness becomes close to 0 and the mean and the median are almost the same: The median is 97% of the mean and it is 3% of a standard deviation from the mean.

To examine how the transformation differentially compresses the data, notice that on the original scale the difference from  $20 \rightarrow 100$  (80 units) is about 4.71 times the difference from  $3 \rightarrow 20$  (17 units), but on the transformed scale the corresponding differences ( $3.00 \rightarrow 4.61$  vs.  $1.39 \rightarrow 3.00$ ) are equal. Transformations for positively skewed data use a  $\lambda < 1$ , thereby compressing the variable more drastically as the variable increases. See Bock (1975), McNeil (1977), Mosteller and Tukey (1977), and Weisberg (1980) for discussion of these transformations.

Returning to Table 1, Cells 3 and 4 indicate different difficulties than found in Cell 2. The multimodality of the data in these cells may indicate that distinct populations, with different distributions, have been combined (see Bradley, 1977). These data require discovering the factor, which may be categorical (e.g., gender; urban vs. rural; with cell phones vs. without cell phones) or continuous (e.g., number of friends) that is found to be associated with the multimodality and statistically removing its effect (e.g., by entering this factor as one or more dummy variables in a regression, if it is categorical, or entering it as a predictor in a regression, if it is continuous). This procedure should presumably create unimodal residuals. This type of transformation is more of a standardization in that the data are made well behaved using the conventional techniques within the general linear model.

### *Transforming to Create Equality of Spread and for Linearizing Relationships*

In addition to transforming data to create symmetry within a single batch, data are also transformed to create *equality of spread* for comparing batches, and for *linearizing relationships* when conducting analyses within the general linear model (e.g., a Pearson bivariate, partial, part, or multiple correlation, regression analysis, or factor analysis). As before, the principal argument here is not about inferential statistics, but about theory and measurement. For example, if the research goal is to describe relationships between variables with a line or curve, the line or curve is a better representation of the relationship if the (vertical) spread around the line or curve were equal across observed levels of the independent variable, which is evidence that the line or curve summarizes the relationship well. Similarly, if for theoretical reasons a linear relationship between independent variables and a dependent variable is hypothesized, it may be necessary to transform one or more variables to attempt to create linearity, and then to test whether the linearity of the relationship is a plausible hypothesis.

Analysis of variance, analysis of covariance, and *t*-tests, as typically practiced, assume homoscedastic residuals. If batches of data are to be compared and they have different degrees of spread, the typical value of each batch is known with different levels of precision, making the comparisons problematic.

If data make the assumption of equality of spread implausible, many transformations could be used to fix this problem besides the one indicated by Equation 1. In addition to the works on exploratory data analysis (e.g., Erickson & Nosanchuk, 1977; McNeil, 1977; Tukey, 1977), most texts on experimental design describe standard transformations to create batches of data with equal spread (e.g., Montgomery, 2009). Focusing on the variance, one measure of spread, there are standard procedures to “stabilize the variance” (i.e., make the variance of the residual unassociated with the level of the independent variable).

The single-bend family of transformations (Equation 1) is especially useful to stabilize the variance when the variances vary as a function of the means of the batches of data. Based on this relationship, Weisberg (1980) indicated the situations in which the square root ( $\lambda = \frac{1}{2}$ ), the logarithm ( $\lambda = 0$ ), and the reciprocal ( $\lambda = -1$ ) will tend to create equality of variance (Table 3, below). Chatterjee and Price (1977, p. 38ff) provide additional variance-stabilizing transformations.

Table 4 shows how one can systematically vary the  $\lambda$  and  $k$  from Equation 1 to find an optimal transformation to create equality of variance. In Table 4 there are two batches of data, and the criterion employed to determine the optimal transformation is the ratio of the larger sample estimate of the population variance to the smaller sample estimate of the population variance. The original (untransformed) data have a ratio of  $66.98:25.17 = 2.66:1$ , but the ideal ratio is 1:1 (i.e., we want these variances to be equal). This ratio is achieved when  $\lambda = 0$  (i.e., a logarithmic transformation) with  $k = -0.5$ . In other words, here  $Y^* = \ln(Y - .5)$ . Furthermore, notice how systematically the ratios in the table change as we vary  $\lambda$  and  $k$ ; this systematic change provides us with a direction for finding an optimal  $\lambda$  and  $k$ . Although in this example

**Table 3** Common Variance Stabilizers

Transformation	Situation	Comments
$\sqrt{Y}$	$\text{var}(e_i) \propto E(Y_i)$	The theoretical basis is for counts from the Poisson distribution
$\sqrt{Y} + \sqrt{Y+1}$	As above	For use when some $Y_i$ 's are zero or very small; this is called the Freeman-Tukey (1950) transformation
$\log Y$	$\text{var}(e_i) \propto [E(Y_i)]^2$	This transformation is very common; it is a good candidate if the range of $Y$ is very broad, say from 1 to several thousand; all $Y_i$ must be strictly positive
$\log(Y+1)$	As above	Used if $Y_i = 0$ for some cases
$1/Y$	$\text{var}(e_i) \propto [E(Y_i)]^4$	Appropriate when responses are "bunched" near zero, but, in markedly decreasing numbers, large responses do occur; e.g., if the response is a latency or response time for a treatment or a drug, some subjects may respond quickly while a few take much longer; the reciprocal transformation changes the scale of time per response to the rate of response, response per unit time; all $Y_i$ must be positive
$1/(Y+1)$	As above	Used if $Y_i = 0$ for some cases
$\sin^{-1}(\sqrt{Y})$	$\text{var}(e_i) \propto E(Y_i)(1-E(Y_i))$	For binomial proportions ( $0 \leq Y_i \leq 1$ )

From *Applied Linear Regression* (p. 124), by S. Weisberg, 1980. New York: John Wiley. Copyright 1980 by John Wiley. Reprinted with permission. Note that " $\propto$ " means "is proportional to" or "varies as."

**Table 4** Example of a Search Space for Finding  $\lambda$  and  $k$  to Equalize the Variance Between Two Batches of Data

$k$	$\lambda$						
	2	1	0.5	0	-0.5	-1	-2
1.0	7.64	2.66	1.72	1.21	1.06	1.24	1.43
0.5	8.16	2.66	1.67	1.15	1.12	1.30	1.46
0.0	8.80	2.66	1.60	1.09	1.18	1.36	1.48
-0.5	9.60	2.66	1.52	1.00	1.28	1.42	1.49
-0.99	10.60	2.66	1.36	1.24	1.47	1.50	1.50

Batch 1: {1.00, 1.00, 1.00, 10.16, 10.16}; sample estimate of population variance = 25.17. Batch 2: {1.00, 1.00, 1.00, 1.00, 19.30}; sample estimate of population variance = 66.98. Within each cell is the ratio of the larger variance to the smaller variance, based on the value of  $\lambda$  and  $k$ . See Equation 1.

we focused on finding one  $\{\lambda, k\}$  combination that works well, there may be others that work equally well, and theory may provide guidance as to the choice of  $\lambda$  and  $k$ .

Another way to create equality of spread for batches of data is to use spread versus level scatterplots as follows: Let  $\ln\text{med}(i)$  = the logarithm of the median for batch  $i$  and  $\ln\text{range}(i)$  = the logarithm of the interquartile range for batch  $i$ . Regress  $\ln\text{range}$  on  $\ln\text{med}$ . Let  $\alpha$  = the slope found from this regression. Assuming that there is a monotonic relation between the medians and the interquartile ranges, the transformation to try to equalize the spread is  $\lambda = 1 - \alpha$  (see Erickson & Nosanchuk, 1977; Montgomery, 2009). Again, note that  $\lambda = 0$  corresponds to a logarithmic transformation.

#### Double-Bend Transformations

Transformations may be used that are *double bend* in appearance (i.e., in the graph of  $Y^*$  on  $Y$ ), typically when the variable to be transformed is *double bound*, that is, bound at the top and bottom with many cases at the floor (minimum) and ceiling (maximum). For example, when data are binomial proportions, the arcsin square root transformation may be used to stabilize the variance of a distribution (see Table 3); in this situation, the variance of the original variable is greater in the middle of the distribution and gets smaller as the variable approaches its endpoints, 1 and 0. To see why this is so, recall that the variance of a binomial variable with probability of success =  $p$  and with  $n$  trials is  $np(1 - p)$ , which has a maximum when  $p = .5$  and approaches 0 as  $p$  approaches 0 or 1.

In addition to the arcsin square root transformation, the single-bend family of transformations may be extended to cases in which a variable is bound at the top and bottom, such as for a test that goes from 0% to 100%. If the data are bunched near the floor (the minimum) and ceiling (the maximum) of the distribution, the data can be stretched by using

$$Y^* = (Y - \text{minimum} + k)^{(\lambda)} - (\text{maximum} - Y + k)^{(\lambda)}, \quad (2)$$

where  $Y^*$ ,  $\lambda$ ,  $Y$ , and  $k$  are defined as in Equation 1. For example, if  $Y$  has the values 0 (minimum), 10, 40, 50, 60, 90, and 100 (maximum), and  $k = 0$  and  $\lambda = \frac{1}{2}$ , then

under this transformation 0 becomes  $-10$ , 10 becomes  $-6.32$ , 40 becomes  $-1.42$ , 50 becomes  $0$ , 60 becomes  $+1.42$ , 90 becomes  $+6.32$ , and 100 becomes  $+10$ . The variable remains symmetric but is stretched more at the tails than at the middle. For additional insight and methods regarding transformations to stabilize variances and to create equality of spread, see Erickson and Nosanchuk (1977, pp. 112–114) and McNeil (1977, chap. 2); McNeil (1977) and Montgomery (2009) also provide techniques for data transformation in factorial designs.

Because of the utility of the general linear model, transformations often are done to linearize relationships that are, in the original variables, nonlinear.<sup>6, 7</sup> The necessity or desire to eliminate the nonlinearity may reflect a strong theory that posits (perhaps by mathematical derivation) that a relationship is linear, or it could reflect finding that assumptions, such as equality of spread, are violated in the nonlinear form and therefore one wishes to create a model that meets these assumptions (for theoretical and descriptive purposes, and not necessarily for inferential purposes). In either case, linearizing the equation is necessary.

Nonlinearity may be observed in the scattergram of the dependent variable on the predictor or in the scattergram of regression residuals on the predicted value of the dependent variable (which, after all, is the linear combination of the independent variables). Weisberg (1980) has shown some common linearizing transformations and the situations to which they apply (see Table 5, below). Notice that these transformations are all within the single-bend family. Additional discussion on transforming data for linearizing relationships may be found in Chatterjee and Price (1977), Fox (1997), Hartwig (1979), Johnson and Bhattacharyya (2006), McNeil (1977, chap. 3), Mosteller and Tukey (1977, chap. 4), and Tufte (1974). Of special note is Mosteller and Tukey's (1977, pp. 84–87) "bulging rule," which assists in

**Table 5** Linearizing Transformations

Transformation		Simple regression form	Multiple regression form
$\log Y$	$\log X$	$Y = \alpha X^\beta$	$Y = \alpha X_1^{\beta_1} X_2^{\beta_2} \cdots X_p^{\beta_p}$
$\log Y$	$X$	$Y = \alpha e^{\beta X}$	$Y = \alpha e^{\sum \beta_j X_j}$
$Y$	$\log X$	$Y = \alpha + \beta(\log X)$	$Y = \alpha + \sum \beta_j \log(X_j)$
$\frac{1}{Y}$	$\frac{1}{X}$	$Y = \frac{X}{\alpha X + \beta}$	$Y = \frac{1}{\alpha + \sum (\beta_j / X_j)}$
$\frac{1}{Y}$	$X$	$Y = \frac{1}{\alpha + \beta X}$	$Y = \frac{1}{\alpha + \sum \beta_j X_j}$
$Y$	$\frac{1}{X}$	$Y = \alpha + \beta \left( \frac{1}{X} \right)$	$Y = \frac{1}{\alpha + \sum \beta_j \left( \frac{1}{X_j} \right)}$

From *Applied Linear Regression* (p. 125), by S. Weisberg, 1980. New York: John Wiley. Copyright 1980 by John Wiley. Reprinted with permission.

determining the  $\lambda$  to employ for both the predictor and the predicted variable within a regression model.

### *Criteria for Transforming Data*

To summarize, transformations are typically used to create meaningful typical values and metrics, equality of spread, and linearity of relationships. However, statistical tests to evaluate these criteria should be subordinate to the idea that one is seeking to find an appropriate metric and functional form of relationship for the variables of interest.

To evaluate the creation of meaningful typical values and metrics, we generally examine a variable's skewness and kurtosis, both in terms of their actual values and relative to their standard errors. The normal distribution provides the standard for comparison: The normal distribution is symmetric (i.e., skewness = 0) and is mesokurtic (neither peaked with fat tails nor flat with thin tails). SPSS (Norušis, 2005) uses the Lilliefors test, which is a modification of the Kolmogorov-Smirnov test, and the Shapiro-Wilk's test to evaluate normality.

There are many statistical tests that may be employed to evaluate equality of spread (see Kirk, 1968, pp. 61–62). Using the variance to evaluate equality of spread across batches, one can create a ratio of the largest sample estimate of the population variance divided by the smallest sample estimate of the population variance, as was done above; this ratio may be tested with Hartley's *F*-max test. Or one can compute the ratio of the largest sample estimate of the population variance divided by the sum of all the sample estimates of the population variance, which is Cochran's *C*; this test requires a Cochran's *C* table. However, both Hartley's *F*-max test and Cochran's *C* are very sensitive to the normality assumption.

There are many other tests that assess equality of spread or equality of variance (e.g., Bartlett's test, which uses the  $\chi^2$  distribution; Levene's test, which uses the *F* distribution). O'Brien's test and the Brown-Forsythe test are less commonly known in the communication discipline but they are more robust than Levene's test and Bartlett's test (see discussion in and sources cited by Algina, Olejnik, & Ocanto, 1989; Olejnik & Algina, 1987; Parra-Frutos, in press).

The SHAZAM program (Whistler et al., 2004) tests the skewness and kurtosis of regression residuals and provides a chi-square test for the normality of residuals that assumes that the residuals come from a homoscedastic population of residuals.

To evaluate linearity, one can examine the data plot or the residuals plot of the variables. SPSS (Norušis, 2005) provides several ways of examining the linearity of the regression model, including tests and graphs regarding the normality and equality of spread of residuals. The normal probability plot (also called a normal scores plot, a quantile-quantile plot, a quantile comparison plot, a Q-Q plot, or a rankit plot; see Bock, 1975; Fox, 1997; Hutcheson & Sofroniou, 2006; Johnson & Bhattacharyya, 2006; Weisberg, 1980) is one way to assess normality of the residuals. In addition, a Box-Cox analysis could be conducted in SHAZAM (Whistler et al., 2004, pp. 163–164), and the estimated  $\lambda$  can be tested against a value of 1, which corresponds to a linear relationship.

Finally, it should be noted that these criteria for data transformation are cumulative: When we compare batches of data (e.g., in ANOVA or in a *t*-test), we would like each batch (or, correspondingly, the residuals in each batch) to be symmetric and unimodal, and, in addition, we would like the batches to have equal spreads. When we conduct a regression, we would like the spread above and below the regression line at each value of *X* to be symmetric and unimodal; the variance at those locations to be equal; and the graph between the residuals and the linear combination of the predictors to be without any discernable pattern.

Although transforming data to achieve one criterion may not provide an optimal transformation based on another criterion, the criteria tend to be optimized jointly; Hartwig (1979) stated that “*nonnormality and nonlinearity often go hand in hand* and, because of this, *reexpression [transformation] is a useful response to both problems*” (p. 54, emphases in original). In addition, some procedures for data transformation use more than one criterion simultaneously. For example, the Box-Cox procedure in SHAZAM (Whistler et al., 2004) seeks to create regression residuals that are both normal and homoscedastic.

### But Aren't Our Inferential Tests Robust Anyway?

Often the response to data transformation is that our inferential tests are robust with regard to violations of their assumptions. The most common assumptions referred to are those that have already been discussed: normality of population errors, homoscedasticity of population errors, and patternless errors. The question of the robustness of inferential tests divides users of statistics into two camps, each with its own goal.

If the data analyst views the goal of the analytic procedures as properly rejecting or failing to reject null hypotheses, even complicated null hypotheses, then data transformation is an unnecessary bother unless the data drastically violate the statistical assumptions. Based on this goal, the idea of transforming data is mostly a nuisance. Unfortunately, this frame of mind does not often lead to a careful examination of the data to determine whether statistical assumptions are met, and the consequences can be substantial. For example, Anscombe (1973) reported four bivariate data sets that have the same means and standard deviations for *X* and *Y*, the same correlation between *X* and *Y*, the same regression slopes, and the same regression intercepts. However, the graphs of the four relationships are remarkably different. As Anscombe stated, “graphs are essential to good statistical analysis” (p. 17). Since Anscombe's article appeared, statistical packages have added statistical indicators and tests of violations of assumptions, but the visual examination of one's data is still essential. Most importantly, unless the value of a statistical model's  $R^2$  or  $\eta^2$  is 100%, the proportion of variance explained does not inform us of the appropriateness of the statistical model employed.

But if the violations of the distributional assumptions are mild and the inferential tests are robust, does one need to transform data? The second camp of analysts would say that even in this case the data need to be examined for possible transformation:

To this camp, the goal is to find the functional form relating the variables and the metric for each variable. For example, Hamblin (1971a) reported rerunning Galileo's inclined plane experiments with the help of his two children, aged 11 and 9. Using relatively crude methods, the proportion of variance explained rounded to 100% using Galileo's equation with appropriate transformation (p. 424). Hamblin commented:

Instead of  $D = AT^2$  [where D is distance traveled, A is a constant reflecting acceleration, and T is time], suppose [Galileo] had blithely assumed that  $D = A + T^2$ . The crucial multiplier effect of the accelerating force would have been missed and his theory would have been grossly inaccurate even though all of the correct variables would have been included. (p. 446)

Hamblin went on to indicate studies in which the wrong functional form (e.g., maintaining a linear form when the evidence indicated that variables should be transformed) reduced the regression  $R^2$ 's by 3% to 50%. However, the levels of those  $R^2$ 's in the original, untransformed variables would be considered large by the standards in most of the social sciences, and the search for a different functional form, one that comes closer to meeting the model's assumptions, would be unlikely to be carried out.

By transforming variables, one is implicitly seeking to find the correct functional form relating the variables in the statistical model. For example, examination of regression residuals for linearity can also provide evidence as to whether a multiplicative effect (an interaction) has been omitted (by observing heteroscedastic residuals), whether a categorical variable with a main effect has been omitted (by observing bimodal or multimodal residuals), and whether a relationship is not linear in the metric of the original variables. In sum, resorting to the robustness of our statistical tests, even when statistically justified, is not a substitute for examining the variables and transforming them when necessary.

### **How Are Forms of Measurement Associated with Types of Data Transformation?**

Some measures are *counts* or *amounts*, which in principle start at 0 and are unbounded at the upper end; in addition, there are derived measures that are based on ratios and differences of counts and amounts. (These ideas are elaborated in Fink, Cai, & Wang, 2006; see also Neuendorf, 2005, pp. 14, 125).<sup>8</sup> These measurement types (counts, amounts, and their derivatives) allow precision, typically evidence high levels of reliability, are the predominant forms of measurement in the sciences, and, more relevant to our discussion, assist in the determination of the functional forms for our variables when assessing our hypotheses. On the other hand, scales that have few response alternatives are imprecise and often make it difficult to find an effective transformation, even if the goal is the modest one of meeting inferential assumptions. This situation is exacerbated when respondents choose few of the scale alternatives. The conclusion is that better measures assist in determining both optimal functional forms of relationships among variables as well as the corresponding data transformations.<sup>9</sup>

The primary measurement characteristics that help in the choice of a transformation are their boundedness and continuity (Table 6). It is common for unimodal single-bound measures, which are likely to be skewed, to be successfully transformed by the single-bend transformation (Equation 1), and double-bound measures that are truncated by the bounds at both ends are likely to be successfully transformed by a double-bend transformation (e.g., Equation 2 and the arcsin square root transformation). When measures have a limited number of response alternatives, such as a scale item with response alternatives {1, 2, 3, 4, 5, 6, 7}, and when only a few of these limited alternatives are actually used, it may be difficult to find an effective transformation.

### What Else is Needed to Determine How to Transform Data?

As in all investigations, the ideal is to have measures that are highly reliable. Bauer and Fink (1983) extended the discussion of data transformation to situations when the variables are measured with nonadditive error and demonstrated how such error affects statistical relationships.

In addition, we need to have enough data points (i.e., a large enough sample) so that it is possible to determine whether and how data should be transformed. Furthermore, the sample should have data over a wide range of values for any noncategorical independent variables.

### Do Scholars Actually Do This? And How Do I Interpret the Results?

Most scientific work uses units that are transformed. A decibel, for example, is based on the unit called a *bel*, and a unit change in bels (i.e., going from  $b$  to  $b+1$  bels)

**Table 6** Characteristics of Measurements with Examples

Boundedness	Continuity	
	Continuous	Discrete
Bound (in principle) at the bottom only*	an amount such as distance or a magnitude scale (Lodge, 1981)	a count such as the frequency of communication between two people
Bound (in principle) at the top and bottom	a mark on a line converted to a number	a score on a <i>true/false</i> test; a Likert scale item
Not bound (in principle) at either the top or bottom	the difference between two amounts, such as change in weight	the difference between two counts, such as change in population

Empirically all scales have a highest value and a lowest value, and therefore are bounded at both the top and bottom. However, in this table what is referred to is the range that the scale could measure, not the results of its application. Furthermore, at the most micro level, all scales appear to be discrete.

\*Scales bound (in principle) only at the top are uncommon; they may be treated as bound at the top by multiplying by -1 or by the subtraction of scale values from the scale's maximum value.

reflects a ten-fold increase in the intensity of a sound. Because of the logarithmic scale used, adding a 60-decibel sound to a 60-decibel sound creates a sound at 63 decibels. There are many other variables defined logarithmically, such as the *bit* (see Absolute Astronomy.com, 2009); these logarithmic variables enter into lawful relationships with other variables.

In the social sciences, there are many instances in which, for theoretical or empirical reasons, transformed variables are used. For example, studies that ask participants to respond to a stimulus may use responses per second or its reciprocal, seconds per response, which indicates the average interval between responses; if the data are reported in one metric (say in responses per second), the reciprocal is a transformation with  $\lambda = -1$ .

Psychophysical research, relating physical stimuli to psychological responses, has a long history of data transformation (Fechner, 1860). Steven's Law proposes that the logarithms of physical and psychological variables are related linearly (discussed with much evidence in Hamblin, 1971a, 1971b). Hamblin (1971a, 1971b) reviewed this literature showing that, with amount (magnitude) measures, Steven's Law worked quite well in relating sociological variables such as wages and liking, education and status, income and status, and, in a multiple regression, vote for Allende for president of Chile in 1952 as a function of several social variables describing the Chilean population. Chatterjee and Price (1977) showed that the regression predicting number of supervisors by number of supervised workers exhibited heteroscedasticity, and a transformation corrected the problem. Tufte (1974) gave several examples of relationships that demonstrate the power of data transformation: He showed that the relationship between parliamentary size and population size and the relation of population size and number of governmental employees were linear when the variables appearing in the regression were logarithmically transformed. In addition, he tested the *cube law*, an equation that relates the proportion of votes for a political party to the proportion of legislative seats won by that party. In this equation the predictor and predicted variable were transformed by a double-bend transformation, namely

$$Y^* = \ln(Y) - \ln(1 - Y), \quad (3)$$

where  $Y$  is a proportion. Furthermore, he reported the nations for which this model's predictions were supported.

The transformation of a variable is specific to the equation or model in which that variable appears. In the absence of both (1) a theory that specifies the functional form of a relationship and (2) measurement rules for the variables that are employed, the transformation used is an empirical matter. In this case the transformation of a variable provides feedback to theory and measurement: It helps determine the subsequent form of the theory and its associated measurement rules.

Are the results using transformed variables interpretable? The analyst needs the theoretical ability to formulate a problem worth studying, skills to measure the relevant variables, and the judgment to design the data collection process. Similarly,

the analyst needs the acumen to examine the data visually and statistically and determine if and how they are to be transformed. If the analyst has all these proficiencies, then the ability to discuss the square root of number of friends or the logarithm of network size should be quite straightforward. And, of course, one can always untransform a transformed variable if that would help create clarity: Square the square-rooted variable, exponentiate the logarithmically transformed variable, etc. The SHAZAM program (Whistler et al., 2004, p. 161) has this feature as an option for its Box-Cox regression routine.

Erickson and Nosanchuk (1977) described the issue this way:

Sometimes there is a sensible interpretation for one of the possible transforms; the choice becomes easier. . . . If a good interpretation is available, from theory or from your own thinking about the data, by all means try the transformation involved. Even if the theoretical transform fails to fit everything perfectly, it will still be useful. . . . On the other hand, if your choice of transformation is based on the data then you'll want to try to come up with an interpretation of your choice. (p. 115)

### **Isn't This Cheating?<sup>10</sup>**

No. On the contrary, analyzing data that fail to meet the inferential assumptions or not extracting any remaining systematic variability in one's variables is cheating: It cheats the community of scholars from understanding what the data could have revealed. Furthermore, in any specific analysis it is likely to be unclear whether transforming data will result in a greater proportion of variance explained or greater likelihood that a null hypothesis will be rejected. Thus, in this limited statistical sense data transformation is not cheating.

### **What Should I Tell My Adviser/Dissertation Committee/Editor?**

There is a great deal of methodological inertia in academic disciplines. Kuhn (1996), for instance, suggested that paradigm changes benefit from the departure—whether by retirement or death or other means—of adherents to the old paradigm, and the same is true for statistical paradigms. Cohen (1990) reminded his readers that “these things take time” (p. 1311), and that W. S. Gosset “published the *t* test a decade before we entered World War I, and the test didn’t get into the psychological statistics textbooks until after World War II” (p. 1311). This view suggests that advisers, dissertation committee members, and editors may oppose or be uncomfortable with the idea of data transformation, regardless of whether the technique is justified by empirical necessity or theoretical derivation. If you understand the role of data transformation in the big picture of developing and testing theories about human communication, and if you take seriously your role as teacher, then it may become your job to teach others about the benefits—descriptive, inferential, and theoretical—of data transformations.

## What Should I Read About This Topic to Be Knowledgeable?

The literature on data transformation is quite extensive and is growing. The references below are annotated to assist the reader in creating a helpful self-taught course. And, within reason, I am willing to discuss these matters with readers.

### Coda

Scientists are detectives, going from theory to measurement to data and back to theory. Transforming data is another step in this investigative process, and it is a valuable step that should not be overlooked: Our job as scientists requires it.

### Notes

- [1] Technically, *residual* refers to sample data whereas *error* refers to population data and, presumably, the true model. Assumptions apply to errors, but assumptions are evaluated by examining residuals. This distinction is important when tests of assumptions are discussed.
- [2] The idea that a variable is “well-behaved” may refer to one or several aspects of a variable. One meaning, which is emphasized here, is that the variable of interest has a relatively symmetric distribution. In other cases the term can refer to a dependent variable that has homoscedastic residuals in a theoretically sensible linear regression.
- [3] Indeed, a variable cannot actually be distributed normally: The tails cannot go to  $\pm\infty$  and actual data cannot be absolutely continuous. However, a variable may approximate a normal distribution.
- [4] The transformation here is both more general and more limited than the standard Box-Cox power transformation (see Box & Cox, 1964; Whistler, White, Wong, & Bates, 2004, p. 155), because we have added a constant but have not used a function that incorporates  $\lambda$  and the geometric mean to keep the units of measurement constant. In addition, if one wished to have the transformation correlate positively with the original scores, one can divide the transformation by  $\lambda$  or multiply the transformation by  $-1$  when  $\lambda$  is negative. See also Bauer and Fink (1983) and Fox (1997, p. 322) regarding this matter. In SHAZAM (Whistler et al., 2004), one may transform (1) the dependent variable only, which is referred to as the classical Box-Cox model; (2) the dependent variable and the independent variables to the same value of  $\lambda$ , which is referred to as the extended Box-Cox model; (3) the independent variables only, each to its own value of  $\lambda$ , which is referred to as the Box-Tidwell model; and (4) all variables, independent and dependent, each to its own value of  $\lambda$ , which is referred to as the combined Box-Cox and Box-Tidwell model.
- [5] The constant  $k$  serves two purposes. First, some values of  $\lambda$  will result in  $Y^*$  being undefined: For example, the logarithm of  $Y$ , if  $Y \leq 0$ , is undefined, as is the square root of a negative number. Thus, if the transformation requires that all  $Y$ s be nonnegative or positive, and some values violate this condition, then a  $k$  can be selected that corrects this problem. Hamblin (1971a, 1971b) associates this constant with correcting for the origin in ratio scales. Mosteller and Tukey (1977) call transformations that employ an additive constant “started” transformations, as in “started logs” and “started roots.” Second, in addition to varying  $\lambda$ ,  $k$  can be varied to search for the optimal single-bend transformation.
- [6] Some nonlinear relations may not be able to be converted to linearity by transforming the original data. Such nonlinear relations are referred to as *intractable*.
- [7] A distinction needs to be made between *linear in parameters* and *linear in variables*. For example, a regression equation is of the form  $\hat{Y} = b_0 + b_1 X_1 + \dots + b_k X_k$ , where  $\hat{Y}$  is the predicted value of the dependent variable,  $b_0$  is the intercept, and  $b_1, \dots, b_k$  are the

coefficients of  $X_1, \dots, X_k$ , respectively. This equation is linear in parameters (the set of coefficients to be estimated). Note, however, that any given independent variable could be a variable that is raised to a power (e.g.,  $X^2$ ), that is an argument of an arithmetic or statistical function (e.g.,  $\log[X]$ ), that is a nonlinear combination of variables (e.g.,  $X_q \times X_p$ ), or of a form other than a variable to the first power. A regression is linear in variables if the variables in the methods regression are variables to the first power. The general linear model is appropriate for equations that are linear in parameters regardless of whether they are linear in variables.

- [8] There are methods to analyze bounded or truncated variables, such as Poisson or negative binomial regression, tobit regression, probit regression, and ordinal logit regression. They may be statistically appropriate alternatives to data transformation (Aldrich & Nelson, 1984; Long, 1997). However, the analyst also needs to consider whether these methods elucidate the interplay of theory and measurement that is fundamental to the discussion in this paper.
- [9] Some of this discussion is taken with little change from Fink et al. (2006).
- [10] Many authors of the literature on data transformation pose this same question for their readers, reflecting in part the social scientist's lack of familiarity and practice with data transformation.

## References

- \*indicates a good source on data transformations for the beginner.  
 \*\*indicates a more advanced, specialized, or difficult source.  
 †indicates a related reading considered important or useful for the general data analyst.  
 ^indicates a general text that includes discussion about data transformation.  
 ~indicates a source not mentioned in the text.
- ~†Abelson, R. P. (1995). *Statistics as principled argument*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Absolute Astronomy.com. (2009). *Logarithmic scale*. Retrieved May 5, 2009, from [http://www.absOLUTEASTRONOMY.COM/TOPICS/LOGARITHMIC\\_SCALE](http://www.absOLUTEASTRONOMY.COM/TOPICS/LOGARITHMIC_SCALE)
- \*\*Aldrich, J. H., & Nelson, F. D. (1984). *Linear probability, logit, and probit models*. Newbury Park, CA: Sage.
- \*\*Algina, J., Olejnik, S., & Ocantos, R. (1989). Type I error rates and power estimates for selected two-sample tests of scale. *Journal of Educational Statistics*, 14, 373–384.
- †Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, 27, 17–21.
- \*\*Bauer, C. L., & Fink, E. L. (1983). Fitting equations with power transformations: Examining variables with error. In R. N. Bostrom (Ed.), *Communication yearbook 7* (pp. 146–199). Beverly Hills, CA: Sage.
- ‘Bock, R. D. (1975). *Multivariate statistical methods in behavioral research*. New York: McGraw-Hill.
- \*\*Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, 26B*, 211–246.
- †Bradley, J. V. (1977). A common situation conducive to bizarre distribution shapes. *The American Statistician*, 31, 147–150.
- ‘Chatterjee, S., & Price, B. (1977). *Regression analysis by example*. New York: Wiley.
- †Cohen, J. (1990). Things I have learned so far. *American Psychologist*, 45, 1304–1312.
- ‘Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- \*Erickson, B. H., & Nosanchuk, T. A. (1977). *Understanding data*. Toronto, Canada: McGraw Hill Ryerson.
- ~\*Fairley, W. B., & Mosteller, F. (Eds.). (1977). *Statistics and public policy*. Reading, MA: Addison-Wesley.

- Fechner, G. T. (1860). *Elemente der psychophysik*. Leipzig, Germany: Breitkopf & Härtel.
- †Fink, E. L., Cai, D. A., & Wang, Q. (2006). Quantitative methods for conflict research, with special reference to culture. In J. G. Oetzel & S. Ting-Toomey (Eds.), *The Sage handbook of conflict communication: Integrating theory, research, and practice* (pp. 33–64). Thousand Oaks, CA: Sage.
- ^Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. Thousand Oaks, CA: Sage.
- ~\*\*Gibbons, J. D., & Stavring, G. R. (1979). Quantitative coefficients for selecting a measure of central location. In K. F. Schuessler (Ed.), *Sociological methodology 1980* (pp. 545–558). San Francisco: Jossey-Bass.
- †Hamblin, R. L. (1971a). Mathematical experimentation and sociological theory: A critical analysis. *Sociometry*, 34, 423–452.
- †Hamblin, R. L. (1971b). Ratio measurement for the social sciences. *Social Forces*, 50, 191–206.
- ~†Hamblin, R. L. (1974). Social attitudes: Magnitude estimation and theory. In H. M. Blalock Jr. (Ed.), *Measurement in the social sciences* (pp. 61–120). Chicago: Aldine.
- \*Hartwig, F. (with Dearing, B. E.). (1979). *Exploratory data analysis*. Newbury Park, CA: Sage.
- ^Hutcheson, G., & Sofroniou, N. (2006). *The multivariate social scientist*. Thousand Oaks, CA: Sage.
- ^Johnson, R. A., & Bhattacharyya, G. K. (2006). *Statistics: Principles and methods* (5th ed.), Hoboken, NJ: Wiley.
- ^Kirk, R. E. (1968). *Experimental design: Procedures for the behavioral sciences*. Belmont, CA: Brooks/Cole.
- \*\*Kruskal, J. B. (1968). Statistical analysis, special problems of transformations of data. In D. L. Sills (Ed.), *International encyclopedia of the social sciences* (Vol. 15, pp. 182–193). New York: Macmillan.
- Kuhn, T. S. (1996). *The structure of scientific revolutions* (3rd ed.), Chicago: University of Chicago Press.
- ~\*Leinhardt, S., & Wasserman, S. S. (1979). Exploratory data analysis: An introduction to selected methods. In K. F. Schuessler (Ed.), *Sociological methodology 1979* (pp. 311–365). San Francisco: Jossey-Bass.
- ~†Lodge, M. (1981). *Magnitude scaling*. Newbury Park, CA: Sage.
- \*\*Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.
- \*McNeil, D. R. (1977). *Interactive data analysis: A practical primer*. New York: Wiley.
- ^Montgomery, D. C. (2009). *Design and analysis of experiments* (7th ed.). Hoboken, NJ: Wiley.
- ^Mosteller, F., Fienberg, S., & Rourke, R. E. K. (1983). *Beginning statistics with data analysis*. Reading, MA: Addison-Wesley.
- \*\*Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression: A second course in statistics*. Reading, MA: Addison-Wesley.
- †Neuendorf, K. A. (2005). *The content analysis guidebook* (6th ed.). Thousand Oaks, CA: Sage.
- ^Norušis, M. J. (2005). *SPSS 13.0 guide to data analysis*. Upper Saddle River, NJ: Prentice Hall.
- \*\*Olejnik, S. F., & Algina, J. (1987). Type I error rates and power estimates of selected parametric and nonparametric tests of scale. *Journal of Educational Statistics*, 12, 45–61.
- \*\*Parra-Frutos, I. (in press). The behaviour of the modified Levene's test when data are not normally distributed. *Computational Statistics* [online]. Retrieved from <http://www.springerlink.com.proxy-um.researchport.umd.edu/content/?k=parra-frutos>
- ~†Shinn, A. M., Jr. (1974). Relations between scales. In H. M. Blalock Jr. (Ed.), *Measurement in the social sciences* (pp. 121–158). Chicago: Aldine.
- ~\*\*Stoto, M. A., & Emerson, J. D. (1983). Power transformations for data analysis. In S. Leinhardt (Ed.), *Sociological methodology 1983–1984* (pp. 126–168). San Francisco: Jossey-Bass.
- †Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: Wiley.
- ~†Torgerson, W. S. (1961). Scaling and test theory. *Annual Review of Psychology*, 12, 51–70.
- \*Tufte, E. R. (1974). *Data analysis for politics and policy*. Englewood Cliffs, NJ: Prentice Hall.

- \*Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.
- ^Weisberg, S. (1980). *Applied linear regression*. New York: Wiley.
- ^Whistler, D., White, K. J., Wong, S. D., & Bates, D. (2004). *SHAZAM econometrics software: User's reference manual version 10*. Vancouver, BC, Canada: Northwest Econometrics.
- †Winer, B. J. (1968). The error. *Psychometrika*, 33, 391–403.
- ~†Woelfel, J., & Fink, E. L. (1980). *The measurement of communication processes: Galileo theory and method*. New York: Academic Press.