DATA TRANSFORMATION

The following brief overview of *Data Transformation* is compiled from Howell (pp. 318-324, 2007) and Tabachnick and Fidell (pp. 86-89, 2007). See the references at the end of this handout for a more complete discussion of data transformation.

Most people find it difficult to accept the idea of transforming data. Tukey (1977) probably had the right idea when he called data transformation calculations "reexpressions" rather than "transformations." A researcher is merely reexpressing what the data have to say in other terms. However, it is important to recognize that conclusions that you draw on transformed data do not always transfer neatly to the original measurements. Grissom (2000) reports that the means of transformed variables can occasionally reverse the difference of means of the original variables. While this is disturbing, and it is important to think about the meaning of what you are doing, but it is not, in itself, a reason to rule out the use of transformations as a viable option.

If you are willing to accept that is it permissible to transform one set of measures into another, then many possibilities become available for modifying the data to fit more closely the underlying assumptions of statistical tests. An added benefit about most of the transformations is that when we transform the data to meet one assumption, we often come closer to meeting other assumptions as well. For example, a square-root transformation may help equate group variances, and because it compresses the upper end of a distribution more than it compresses the lower end, it may also have the effect of making positively skewed distributions more nearly normal in shape. If you decide to transform, it is important to check that the variable is normally or nearly normally distributed after transformation. That is, make sure it worked.

When it comes to reporting our data… although it is legitimate and proper to run a statistical test, such as the one-way analysis of variance, on the transformed values, we often report means in the unit of the untransformed scale. This is especially true when the original units are intrinsically meaningful. Howell (2007) urges researchers to look at both the converted (transformed) and unconverted (original) means and make sure that they are telling the same basic story. Do not convert standard deviations – you will do serious injustice if you try that. And be sure to indicate to your readers what you have done. It is not uncommon to see both the converted and unconverted values reported. Tabachnick and Fidell (2007) point out that, although data transformations are recommended as a remedy for outliers and for failures of normality, linearity, and homoscedasticity, they are not universally recommended. The reason is that an analysis is interpreted from the variables that are in it, and transformed variables are sometimes harder to interpret.

You should not get the impression that data transformations should be applied routinely to all your data. As a rule of thumb, "If it's not broken, don't fix it." If your data are reasonably distributed (i.e., are more or less symmetrical and have few, if any, outliers) and if your variances are reasonably homogeneous, there is probably nothing to be gained by applying a transformation. If you have markedly skewed data or heterogeneous variances, however, some form of data transformation may be useful. Furthermore, it is perfectly legitimate to shop around for a transformation that makes the necessary changes to the variance and shape. The only thing you should not do it to try out every transformation, looking for one that gives you a significant result. You are trying to optimize the data, not the resulting *F*.

DATA TRANSFORMATION

As suggested by Tabachnick and Fidell (2007) and Howell (2007), the following guidelines (including SPSS compute commands) should be used when transforming data.

If your data distribution is…	Use this transformation method.
Moderately positive skewness	Square-Root NEWX = SQRT(X)
Substantially positive skewness	Logarithmic (Log 10) NEWX = LG10(X)
Substantially positive skewness (with zero values)	Logarithmic (Log 10) NEWX = LG10(X + C)
Moderately negative skewness	Square-Root NEWX = SQRT(K – X)
Substantially negative skewness	Logarithmic (Log 10) NEWX = LG10(K – X)

C = a constant added to each score so that the smallest score is 1.
K = a constant from which each score is subtracted so that the smallest score is 1; usually equal to the largest score + 1.

References

Howell, D. C. (2007). *Statistical methods for psychology* (6th ed.). Belmont, CA: Thomson Wadsworth.

Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology, 68*, 155-165.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Allyn and Bacon.

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

Also see:

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1983). *Understanding robust and exploratory data analysis*. New York: Wiley.