Examine distributions and relationships

To spot data quality problems and begin to look for associations among variables, perform univariate and bivariate analysis.

A. Univariate analysis. Properly examining, cleaning and transforming the data are crucial and must be done before analysis. To examine the distribution of values for each variable:

i) If the variable is categorical, compute frequencies of response values and generate bar charts for easy examination. Are there at least a few values in each category? This is important because there first needs to be a minimum amount of variation in the data to assess whether program participation (or some other factor) is associated with change.

ii) If the variable is continuous, compute descriptive statistics (mean, median, standard deviation, range) and generate a histogram and a boxplot for easy examination. Is the distribution of a continuous variable "multi-modal", with two or more peaks, or skewed, with many values at one end of the distribution and few at the other? If so, it may first be necessary to transform the data prior to summarizing or drawing conclusions from it. Is the distribution of a continuous variable roughly bell-shaped (aka "normally distributed"), with many responses sitting in the middle, and a few below and a few above? If so, no transformation is necessary for descriptive statistics to be meaningful summaries of the data, and statistical tests which rely on an assumption of normality may be used.

On transformations: if a very small number of or no response values in a response category, combine two "emptier" categories into a broader, "fuller" one (e.g. age 9-11 and age 12-14 combine to form age 9-14.); b) if response values are not normally distributed, either apply a transformation (most commonly log or square root transformations) or recode a continuous variable into a categorical one (e.g. recode individual age values into age range groups such as the one above); c) if comparing a pre- to a post-score, compute a "difference score" or "change score" by subtracting pre-score from post-score.

Make sure to add a description for each new variable to the codebook.

Then perform univariate analysis on all transformed variables, and examine their distributions.

iii) Look for surprising values. If found, assess the reasons and make necessary adjustments, either to data or interpretation.

iv) Look for missing values. Make sure they are properly entered according to your missing value codes. If there is a large number of missing values for a particular item, make appropriate adjustments, either to data (transform or remove that item) or to interpretation. Consider revising question wording or data collection procedures for the

next administration of the measure. Also, note how many completed records there are for each variable.

Once the data has been cleaned, examine the distribution for each variable. For some evaluations, descriptive statistics and frequency distributions of individual variables are all that is needed (see Ellen Taylor Powell "Analyzing Quantitative Data"). However, if the evaluation questions ask whether the program is *associated with or causes change* in participant outcomes, or asks whether certain factors are associated with different levels of change, it will be necessary to analyze the relationships among pairs, and sometimes groups of variables.

B. Bivariate analysis. To spot data quality problems and begin to look for associations among variables, check to see whether the response values for each possible variable pair appear to move together. Does the strength of the association for each variable pair appear as expected? Unexpectedly strong (or weak) associations may indicate a problem with question wording, data collection or data entry. To check variable pairs depends on the level of measurement of each variable:

i) If two categorical variables, compute cross tabulations in a contingency table and ask for row and column percentages. Look for unusual patterns in the percentages.

ii) If two continuous variables, compute the correlation coefficient. Look for high correlation coefficients (close to -1.0 or 1.0)*.

iii) If one categorical and one continuous variable, for each response category of the categorical variable, compute the mean of the continuous variable, and compare. Look for striking variation in the mean of the continuous variable across different categories.

*What is considered a high value varies. Consult the evaluation literature on similar programs for guidelines.