Steps For Data Collection and Quantitative Analysis

Taking the time to properly plan data collection and analysis yields better quality data and more credible results. The following is an introduction to working with alphanumerically coded data (e.g. multiple choice surveys, knowledge tests). This is meant as a general guide only. When thinking through choices about how to collect, analyze and present data, always keep in mind the logic of the specific evaluation question being investigated, the lifecycle stage of the program and stakeholder needs. This material includes: 1) preparing for data collection; 2) collecting data and 3) examining results 4) testing for statistical significance and 5) attributing participant change to your program. Terms highlighted in red are defined in a glossary at the end. References to other resources follow.

Two cautions: First, measures with standardized response options are neither the only, nor are they always the best way to gather evidence. These lend themselves easily to aggregation across large numbers of participants, and quantitative analysis. However, depending on the evaluation question, qualitative measures and qualitative analysis are often called for, especially where richer, descriptive information is needed to explain complex dynamics or developmental processes, and the sample is smaller. Sometimes, quantitative and qualitative approaches are both used to answer the same evaluation question.

1) Prepare for data collection.

A. **Generate a list of variables**. For each EQ and its accompanying comparison strategy, decide on the unit of observation and make a list of the variables on which it will be necessary to collect data, as follows:

Unit of observation--the thing on which data is collected. This may be a person, a parent/child pair, an organization, a program site, a region, etc. Before starting to collect data, make a list of all of the possible units of observation on which it might be useful to compare results. *For example*, it might be useful to individually identify participants (whether anonymously or by name) *and* to note the program site at which the participant engaged in the program.

V*ariable*--a characteristic(s) of the unit of observation. Before starting to collect data, look at the EQs and comparison strategies, and make a list of variables on which data should be collected. *For example*, the EQ might be, "Is program participation associated with a change in score?" Here the variables are program participation and score. However, you might also want to know if there is a difference in how much male v. female participants change and in how younger v. older participants change. Here, there would need to be variables for gender and age as well.

B. C**reate a spreadsheet**. For smaller databases, this could be done in Excel or Microsoft Access. For larger databases, or for easier analysis of small databases, this could be SPSS, Jump, or Minitab. Licenses can often be obtained through your university. Organize the spreadsheet using the units of observation and variables chosen above, with each variable listed as a column heading and a participant's name or anonymous person identifier code leading off each row. Any group identifier code (such as one to denote program site) can be entered in a separate column (see example p.3).

C**. Create a codebook.** Data must be abbreviated (or coded) for inputting and analysis. A codebook translates richer information into abbreviations, and tells location of each piece of data in the set. To create a codebook, for each variable, list the following:

i). Variable name: start with a letter; use eight characters or fewer; use no spaces or symbols (underscore is okay);

ii). Variable label or description: explain what the variable name stands for; give the unit of measurement;

iii). Level of measurement: refers to a question's response options; may be either qualitative/categorical (ordinal or nominal); quantitative/continuous (interval or ratio); or Likert (may be either qualitative or quantitative, depending);

iv). Value labels: explains what each response value code stands for, including codes for missing values;

v). Location of data for each variable within the larger data file (note that in statistical software programs, each character or numeral occupies one column--if using Excel, this does not necessarily apply).

vi). Formats: tells how many characters or numerals; number of decimal places.

Example codebook:

Variable name	**Variable label or description**	**Level of measurement**	**Value labels**	**Location**	**Format**
PersonID	Participant constructed, unique identifier	n/a	n/a	col.1-8	XXXXXXXX
ProgSite	Name of site where participant engaged in program	categorical	.=missing 1=campus 2= Extension office 3=high school	col.9	X
Age	Age in years at last birthday by self report	continuous	.= missing value = age in years	col.10-12	XXX
Gender	Gender by self report	categorical	.=missing 1=male 2=female	col.13	X
ProgPart	Program group or comparison group	categorical	. =missing 1=…participant 2=…comparison	col.14	X
PreScore	Score on pre-test	continuous	.=missing range=0-100	col.15-17	XXX
PostScore	Score on post-test	continuous	.=missing range=0-100	col.18-20	XXX

2) Collect data. To assure consistency across data collectors and across time, clearly articulate and pilot test data collection procedures for each EQ. Here is a list of steps:

A. Decide how long to spend collecting and analyzing data then set aside time and resources for data collection, entry, and analysis.
B. Write data collection procedures.
C. Pilot test measures. Observe data collection procedures and examine data.
D. Revise measures, guides, and timeline as needed.
E. Train data collectors so that data are collected in a consistent way.
F. Have data collectors document their actual data collection procedures, including any modifications to procedure.
G. Perform data entry and examine data on an ongoing basis. Adjust data collection procedures as needed.

Reminders: Back up electronic data frequently. Keep hard copies of data in a well-known location, to check data entry and in case electronic files are lost.

Example spreadsheet (imagine 20+ Person IDs):

Person ID	**Program Site**	**program participation**	**age**	**gender**	**pre score**	**post score**
0004JE	1	1	16	1	70	80
5014NE	3	1	12	1	80	80
1008AE	2	1	15	1	75	70

Once you have collected data, don't let the data get "stale" on the shelf. As soon as possible after data collection is complete, start preparing to analyze it. If possible, analyze data as you are collecting it, so that problems with measures or data collection procedures become known sooner and can be corrected early on.

3) 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"). This may be all you need to do, or, if the data has serious flaws, it may be all you can legitimately do. 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.

4) Test statistical significance of distributions and relationships. If two variables of interest appear correlated, claims that the observed relationship has a connection to program theory or is generalizable to a larger population are not valid unless it can be shown, using statistical tests, that the observed relationship is unlikely to have occurred due to other factors, or by random chance. Statistical tests may be used to rule out chance as an alternate reason for observed relationships among variables. The table below presents, in the rightmost column, some possible statistical tests, with most commonly used tests in bold type. The appropriate statistical test to use depends on the evaluation question to be answered and type of relationships being investigated, the number of variables and the level of measurement of variables in the analysis. However, remember that many of these tests rely on specific assumptions about the data, such as sample size and the shape of the distribution of data on each variable, which must be checked. Before applying tests of the statistical significance of observed relationships among variables, look up the test in a statistics text or authoritative online resource to make sure the data satisfies all necessary assumptions (see "Sources" listed at end).

When data don't have a robust distribution (N>30, bell-shaped distribution), familiar, "parametric" statistical methods such as a t-test lose their power and are no longer appropriate. In this case, DO NOT GIVE UP. Non-parametric methods will often still work! Non-parametric statistics are tests of statistical inference that **don't** rely on assumptions of parametric statistics such as, e.g. that the data are randomly drawn from a normally distributed population, have few influential outliers, and that N>30. Non-parametrics can be used with N as small as 5, with data that is skewed, or has unequal sample sizes (see Siegal and Castellan 1988, Gibbons 1993).

5) Attributing participant change to your program. Many evaluation questions ask about "effectiveness" of a program. Even if you can show participants changed on a target outcome and the change is statistically significant, to make a strong claim of "effectiveness", you must rule out rival explanations external to your program, such as contextual factors. To elaborate on the relationship between the program variable and the target variable, divide or "stratify" your data by a relevant contextual variable (e.g. males and females; low education and high; participants and non-participants in a competing program) to see whether results seem to depend on being in a particular sub-group. Then examine the relationship between the program participation and the outcome variable for each sub-group. (see Isreal 1992)

Even if all conditions for strong claims are met, always report results along with a description of any limitations of the data. If, after collecting data, it is unclear whether the results can support claims of effectiveness, consult a statistician.

Univariate, bivariate and multivariate analysis choices

*(*Items in **bold** are most commonly used*)*

Identify variables of interest		***Examine data distributions and relationships***		***Test data distributions and relationships***
Number of variables	*Type(s) of variables*	*Statistical summary*	*Graphical summary*	*Possible tests of significance (note: not an exhaustive list!)*
1	1 continuous variable	Descriptive statistics	histogram, boxplot	**1-sample t-test**
	1 categorical variable	Frequencies	bar chart	**chi-square goodness of fit**
2	2 categorical variables	Cross tabulations	cluster bar chart	**-[Pearson]chi-square test of independence** -Fisher's Exact test -McNemar's test
	1 continuous and 1 categorical variable	Compare means	bar chart, histogram by panel, box plots	**-paired samples t-test** (if matched pre/post or if matched T and C) **-independent samples t-test** (if randomly selected T and C) -analysis of variance (ANOVA) -regression analysis -Mann-Whitney U test -Wilcox Rank Sum test -Kruskal-Wallis test -Friedman test
	2 continuous variables	Compute correlation coefficient	scatterplot	-**Pearson's Product Moment r** -Spearman correlation -Regression analysis
3+	3+ variables	split the file by one of the variables and revert to two bivariate analyses OR compute multiple regression and partial correlation coefficients	3-D scatterplot; clustered barchart	**multiple regression analysis** -factor analysis -cluster analysis -multilevel models -logistic regression

Glossary

qualitative analysis-the non-numerical examination and interpretation of observations for the purpose of discovering underlying meanings and patterns of relationships. (R&B)
quantitative analysis-the numerical representation and manipulation of observations for the purpose of describing and explaining the phenomena that those observations reflect. (R&B)
spreadsheet-a document that organizes data in rows and columns of cells
codebook-the document used in data processing and analysis that tells the location of different data items in a data file. Typically, identifies the locations of data items and meaning of the codes used to represent different attributes of variables (R&B)

Level of measurement-refers to the relationship among the values that are assigned to the attributes for a variable. (T)

Qualitative/Categorical:
nominal-values assigned to the attributes for a variable are just placeholders for longer text items; there is no assumed ordering among the values (e.g. variable is party membership and values are 1=Republican, 2=Democrat 3=Independent). (T)
ordinal-values assigned to the attributes for a variable are in rank order, but distances between values are not meaningful. (E.g. variable is education attainment and values are 1=less than HS diploma, 2=HS diploma, 3=some college, 4=4 year degree.) (T) However, if the distances between values are meaningful, an ordinal variable may be considered continuous. (FV)

Quantitative/Continuous:
interval-values assigned to the attributes for a variable are in order and distances between each value and the next are constant, interpretable; averages and other statistics may be meaningfully computed across observations. (T)
ratio-like interval, but the set of values assigned to attributes include an absolute zero; fractions may be meaningfully computed.(T)

Likert scale-a type of composite measure developed by R. Likert in an attempt to improve the levels of measurement in social research through the use of standardized response categories in survey questionnaires. Likert items are those using such response categories as "strongly agree", "agree", "disagree", and "strongly disagree". (R&B) May be considered either categorical or continuous, depending on whether distances between response values are meaningful. (FV)

Univariate analysis-the examination of the distribution of units of observation (participants, groups) on only one variable at a time. (R&B)
descriptive statistics-statistical computations describing either the characteristics of a sample or the relationship among variables in a sample; summarize a set of sample observations; include mean, median, mode, variance, standard deviation and range. (R&B)
frequency distribution-a description of the number of times the various attributes of a variable are observed in a sample. For example: "53% of the sample were men and 47% were women". (R&B)
statistical inference-from findings based on sample observations, drawing conclusions about some larger population or about the theoretical meaning of observed relationships among variables. (R&B)

Glossary

Bivariate analysis-the examination of the joint distributions of units of observation (participants, groups) on two variables at a time in order to assess the independence or non-independence of the two distributions.

statistical significance-a general term referring to the unlikeliness that relationships observed in a sample could be attributed to sampling error alone; a relationship between variables is said to be statistically significant when its probability of occurring due to chance is at or below a cutoff point selected in advance; claiming statistical significance for such a relationship means it can be generalized beyond the sample and reflects more than chance covariation. (R&B)

cross tabulations- the process of creating a contingency table from the frequency distributions of the response data for statistical variables

contingency table-a table format for presenting the relationships among variables—in the form of percentage distributions. Shows values of the dependent variable contingent on values of the independent variable. (R&B)

correlation-a single number that describes the degree of relationship between two variables (T)

regression analysis-a general statistical analysis that enables us to model relationships in data and test for treatment effects; models relationships that can be depicted in graphic form with regression lines. (T)

In-text citation key:

"FV"= Francoise Vermeylen

"R&B"=Rubin and Babbie "Research Methods for Social Work"

"T"=Trochim Research Methods Knowledge Base

Sources:

Much content is adapted from the "Basic Data Management" workshop presented by Francoise Vermeylen, Cornell Statistical Consulting Unit (FV)at Mann Library, Cornell University campus in the fall of 2009. For more from CUSCU, see: http://www.cscu.cornell.edu

For resources on research methods, generally, see:
Trochim, W.M.K. "The Research Methods Knowledge Base" (T) http://www.socialresearchmethods.net/kb/
Rubin, Allen and E. Babbie (1997) (R&B) Research Methods for Social Work, Cole Publishing Company,
especially Chapter 15, "Interpreting Descriptive Statistics and Tables" and Chapter 16, "Inferential Data Analysis".

For more information on generating descriptive statistics, rankings and cross tabulations, see especially:
Powell, Ellen-Taylor (1996)"Analyzing quantitative data"
University of Wisconsin Extension, publication G-3658-6 http://learningstore.uwex.edu/Assets/pdfs/G3658-06.pdf

For more information on collecting evaluation data, see University of Wisconsin Extension Evaluation Publications:
"Collecting evaluation data; an overview of sources and methods" publication G3658-4 and others at
(http://www.uwex.edu/ces/pdande/evaluation/evaldocs.html):

For more information on performing simple analyses on survey data in Excel, see:
Leahy, Jennifer (2004) "Using Excel for analyzing survey questionnaires"
University of Wisconsin Extension, publication G3658-14 at http://learningstore.uwex.edu/Assets/pdfs/G3658-14.pdf

For more information on presenting evaluation results in graphic form, see:
Minter, Ed and Michaud, M. (2003) "Using graphics to report evaluation results"
University of Wisconsin Extension publication G3658-13 at http://learningstore.uwex.edu/%2FUsing-Graphics-to-Report-Evaluation-Results-P1022.aspx

For more information on non-parametric statistics, see:
Nonparametric Statistics for the Behavioral Sciences", by Siegel and Castellan (1988)
Non parametric statistics, an introduction, (Gibbons 1993) Sage Series in Quantitative Applications #90

For more information elaborating program impacts, see:
Israel, Glenn D. 1992. Elaborating Program Impacts Through Data Analysis. Program Evaluation and Organizational Development, IFAS, University of Florida. PEOD-3, September. http://edis.ifas.ufl.edu/pd003