

Data Screening Check List

Refer to *Using Multivariate Statistics* (Tabachnick & Fidell, 2007, 5th Ed).

Name of Data file _____

Circle one: Ungrouped Data Grouped Data

Note: The order for grouped data (e.g., blessings vs. hassles vs. controls) is the same except that screening for homogeneity of variance is added to Step 5B. Also, much of the screening is done within groups instead of within the whole sample.

Step 1. Data Reentry

A personal communication with Dr. William Chaplin (February, 5, 2008, St. John's University, NY)

“My view is that errors in data entry contribute (probably random) variance to the data. In the behavioral sciences we are generally studying phenomena that is so complex that sorting out the effect in the context of random and systematic (i.e. unmeasured variables) is difficult enough without adding more noise to the system. Said another way, inaccurately entered data hurts power and we can reduce inaccurate data through double entry....

[If you have a large data set or other people on the project want to avoid double entry]...one solution is to spot check your data and see if the error rate is acceptably low (THERE IS NO RULE for acceptable). To me 0% is acceptable (see above), to others maybe 5% and only double enter if the error rate is unacceptable (e.g., 25%).

One could also consider who is doing the entry. If the PI is entering his or her own data it is probably more accurate than if it is some undergraduate or secretarial staff.

But my position is that there are many things that we cannot control about the accuracy of our data—one thing we can control is accuracy of data entry—so we should take advantage of that.”

I agree with Dr. Chaplin 100%!

SPSS has a nice comparison program in a module called “Data Entry Builder.” Here’s the link:

http://www.spss.com/Data_Entry/data_entry_builder/

Another approach (if the dataset is small) is to have someone, manually and individually, compare the raw data with the entered data and make any corrections along the way.

- i. Were the data reentered? Yes No
- ii. If yes, how? Data Entry Builder Manually
- iii. If yes, by whom? _____

- iv. If no, conduct a spot check.
- v. Based on the spot-check, what percentage of data were entered inaccurately (i.e., the error rate)? _____ %

Date Verified _____

Verifier Name (please print) _____

Step 2. Inspect Univariate Descriptive Statistics for Accuracy of Input

A. Out-of-range values (DO THIS FOR EVERY INDIVIDUAL ITEM SEPARATELY): Were the data entered correctly? If possible the data should be proofread against the original data (on the questionnaires, etc.) to check that items have been entered correctly. Preferably someone other than the person who entered the data should do this. Another option is for the data to be entered twice and superimposed onto each other. A perfect fit suggests that data have been entered identically on both occasions. These methods are highly recommended with small datasets. Because they're inefficient and perhaps impossible for larger datasets, you can do this via SPSS FREQUENCIES.

- i. For **continuous variables**, are all the values within range? Yes No

If no, complete the table below:

What variable(s)?	Case Number	Old Value	New Value	How handled?

Date Verified _____

Verifier Name (please print) _____

ii. For **dichotomous variables** (sex, yes/no), are all the values within range?

Yes No

If no, complete the table below:

What variable(s)?	Case Number	Old Value	New Value	How handled?

Date Verified _____

Verifier Name (please print) _____

iii. For **discrete variables** (ethnicity, categories of religious affiliation), are all the values within range?

Y N

If no, complete the table below:

What variable(s)?	Case Number	Old Value	New Value	How handled?

Date Verified _____

Verifier Name (please print) _____

iv. Are the **codes for missing values** (i.e., 999 as a discrete variable) accurately programmed for all variables? *(In SPSS 14 variable view click on the missing values box and select discrete missing values, in the first box enter 999, click on OK).*

Yes No

If no, complete the table below:

What variable(s)?	Case Number	How handled?

Date Verified _____

Verifier Name (please print) _____

Create Composite Variables for Step 2b and Step 2c AFTER YOU HAVE PROPERLY TRANSFORMED ALL REVERSED ITEMS FOR YOUR SCALES

What composite variables were created?	What items were used to create this composite?	Which items, if any, were reverse scored?	How did you create the composite? (mean, sum)

--	--	--	--

Note: After creating the syntax for the composite variables (in SPSS, TRANSFORM/compute/paste), what did you name the file?

Date Verified _____

Verifier Name (please print) _____

B. Plausible means and standard deviations:

Are the means and standard deviations plausible? Yes No

If no, complete the table below:

What variable(s)?	Actual <u>mean</u>	Possible reason for being an implausible <u>mean</u> .	Actual <u>Standard Deviation</u>	Possible reason for being an implausible <u>standard deviation</u> .	How handled?

Date Verified _____

Verifier Name (please print) _____

C. Univariate outliers: These are case scores that are extreme and therefore have a much higher impact on the outcome of any statistical analysis. **At this point, you only identify the univariate outliers as a means of determining if the data were inputted accurately. They're dealt with in Step 5A.**

a. Four basic reasons you'd get an outlier:

- i. There was a mistake in data entry (*a 5 was entered as 55, etc.*). Hopefully step 1A above caught all of these.
- ii. The missing values code was not specified and missing values are being read as case entries (*999 is typically the missing value code but it will be read as an entry and possibly an outlier if the computer is not told that these are missing values*). Hopefully Step 1A above caught all of these.
- iii. The outlier is not part of the population from which you intended to sample (*you wanted a sample of 7 year olds and the outlier is a 10 year old*). In this case the outlier should be removed from the sample.
- iv. The outlier is part of the population you wanted but in the distribution it's seen as an extreme case.

b. Detecting outliers:

- i. Among **dichotomous variables:** If you have a dichotomous variable with an extremely uneven split (i.e., 90 – 10 split, 90% say “yes” and 10% say “no”) this will produce an outlier. This is identified via SPSS FREQUENCIES. The most common fix for this is to delete the variable. You can also retain the variable, but must realize that its association with other variables will be deflated.

What dichotomous variables had uneven splits?	Did you delete them? Yes or No?	If no, why?

Date Verified _____

Verifier Name (please print) _____

- ii. Among **continuous variables**: Whether searching for univariate or multivariate outliers the method depends on whether the data are grouped or ungrouped. If you're performing analyses with ungrouped data (e.g., regression, canonical correlation, factor analysis, or structural equations modeling) univariate and multivariate outliers are sought among all cases. If you're going to perform the analyses with grouped data (ANOVA, ANCOVA, MANOVA, MANCOVA, profile analysis, discriminant function

analysis, or logistic regression) both univariate and multivariate outliers are sought within each group separately.

Univariate outliers are those with very large standardized scores (z scores greater than 3.3) and that are disconnected from the distribution. SPSS DESCRIPTIVES will give you the z scores for every case if you select *save standardized values as variables* (the z scores are saved in the data file). As an alternative, or in addition to inspection of z scores, there are graphical methods for finding univariate outliers. SPSS FREQUENCIES will give you histograms (use SPLIT FILE/ Compare Groups under DATA for grouped data). Boxplots are simpler and literally box in observations that are around the median; cases that fall far away from the box are extreme. **I prefer the z score approach (adding Boxplots is sometimes helpful).**

List all univariate outliers here and how they were handled.

What variable(s)?	Case Number	Z score	Reason for being an outlier (refer to Step 1C a).	How handled (correct data entered, discrete variable created, deleted). If the reason is #4 in Step 1C a, you'll treat it in Step 5A.

--	--	--	--	--

Date Verified _____

Verifier Name (please print) _____

Step 3. Evaluate Amount and Distribution of Missing Data; Deal with Problem

“The important thing in dealing with missing data is to figure out if the data are missing randomly (Missing Completely at Random--MCAR or Missing at Random--MAR) or if there is some pattern (reason) to why the data points are missing (Missing Not At Random--MNAR). If only about 5% or less of the data are MCAR or MAR from a large data set, almost anything thing you do will yield similar results. Unfortunately, there are no firm guidelines for how much missing data can be tolerated for a sample of a given size” (Tabachnick & Fidell).

EQS: If you have access to EQS, you should run EQS’ missing data diagnosis, available via ANALYSIS/Missing Data Analysis. This output provides information about missingness on individual variables and pairs of variables.

“As part of this output, EQS also provides a simple correlational diagnosis that can help to evaluate the pure randomness of missing information, a special case of MCAR. Under pure randomness, you should observe low correlations because missingness or presence of data on one variable should not be predictable by presence or absence of data on another variable” (Bentler, 2006).

Note: The above is not an actual test of MCAR, which is available in EQS when modeling (SEM).

Circle one: What’s the pattern of data? MCAR MAR MNAR

How did you make this determination?

Date Verified _____

Verifier Name (please print) _____

“Create a dummy variable that keeps track of the missing values so it can be used as a variable later (Differences between complete and incomplete subjects, etc). So you should create a variable for cases with complete data (coded as “0”) and cases with missing data (coded as “1”)” (Tabachnick & Fidell, 2001).

Did you do this? Yes No

Date Verified _____

Verifier Name (please print) _____

Circle one: How did you impute missing data?

Estimation Maximization (using EQS) or Multiple Imputation (using NORM)

Note: We should repeat analyses with and without missing data. This is particularly important if the data set is small, the proportion of missing data is high, or data are MNAR. “If the results are similar, you can have confidence in them. If they’re different, however, you must investigate the reasons for the difference, and either evaluate which result better approximates “reality” or report both sets of results” (Tabachnick & Fidell, 2001).

Step 1: Because I’m ultra conservative with treating missing data, I first focus on treating missing data for individual items. For example, say there is five items that reflect different positive affect adjectives (i.e., gratitude, hope, love, joy, happy). Say a participant gave responses for all adjectives except “happy.” Using EM in EQS, I would then predict that participant’s “happy” score with the other four affect adjectives.

What item(s) had missing	What items were used to	Case	Did you do a spot check on
--------------------------	-------------------------	------	----------------------------

values?	impute the missing value (e.g., love and joy were used to impute values for happy)?	Numbers	the imputed values to ensure they seem probable? For example, if happy ranges from 1-6, is the imputed value 3.36 (which is probable)? Or 12.13 (which is improbable)? Yes or No

--	--	--	--

Date Verified _____

Verifier Name (please print) _____

CALCULATE THE ALPHA FOR ALL SCALES. IN SPSS, ANALYZE/scale/reliability analysis. You should only consider analyzing scales with an alpha that's $\geq .70$.

NOW CREATE NEW COMPOSITE VARIABLES (ideally only for scales with an alpha that's $\geq .70$) BECAUSE MISSING DATA HAVE BEEN FIXED WHERE NEEDED FOR ALL INDIVIDUAL ITEMS. USE THESE NEW COMPOSITE VARIABLE FROM THIS POINT FORWARD.

Step 2: Although you've now treated missing data for all individual items, you may still be missing data for some composite variables because a participant failed to complete *all* of the items for the specific scale. Using the example in Step 1 of treating missing data, if a participant fails to provide responses for all of the five positive affect adjectives, it's impossible to predict scores for individual items. Therefore, here in Step 2, you use other composites where you have data to predict composites where you have no data. For example, say a participant fails to complete any of items for the five positive affect adjectives, thus leaving you with no "positive affect" composite. But they thankfully provided enough data on other items so that you have a "life satisfaction" composite and "gratitude" composite. Using EM in EQS, I would then predict that participant's "positive affect" composite with their "life satisfaction" composite and "gratitude" composite.

What variable(s) had missing values?	What variables were used to impute the missing value (e.g., life satisfaction and gratitude were used to impute values for positive affect)?	Case Numbers	Did you do a spot check on the imputed values to ensure they seem probable? For example, if positive affect ranges from 4-20, is the imputed value 11.26 (which
--------------------------------------	--	--------------	---

			is probable)? Or 27.60 (which is improbable)? Yes or No

--	--	--	--

Date Verified _____

Verifier Name (please print) _____

Step 4. Check Pairwise Plots for Nonlinearity and Heteroscedasticity

Nonlinearity: “The assumption of linearity is that there is a straight-line between two variables (where one or both of the variables can be combinations of several variables)” (Tabachnick & Fidell). Conduct bivariate scatterplots between pairs of variables. If both variables are normally distributed and linearly related, the scatterplot is oval shaped. If one of the variables is nonnormal, then the scatterplot between this variable and the other isn’t oval. Go to GRAPHS/scatterplot/simple/define and put one variable in the Y-axis and the other variable in the X-axis. If the scatterplot isn’t oval, you might have to transform the variable(s) that have high skewness or kurtosis (see Step 5B). Here are the three most common transformations, from weakest to strongest: square root [SQRT(your variable)], logarithmic transformation [LG10(your variable)], negative reciprocal [-1000/your variable]. **Note:** If a variable is negatively skewed, it first must be reflected by subtracting every score from a constant that is one greater than the highest score. So if the highest score for the variable “gratitude” is 100, in SPSS go to TRANSFORM/compute and create a variable called “gratitude_ref” by putting the following in the Numeric Expression box: 101-gratitude. Click OK.

“Conducting bivariate scatterplots makes sense if there are only a few variables. But if there are numerous variables, you should use statistics on skewness to screen only pairs that are likely to depart from linearity. Think, also, about pairs of variables that might have true nonlinearity and examine them through bivariate scatterplots” (Tabachnick & Fidell).

Heteroscedasticity (the failure of homoscedasticity): This is an assumption for analyses using **ungrouped univariate data**. Variables are homoscedastic when “the variability in scores for one continuous variable is roughly the same at all values of another continuous variable” (Tabachnick & Fidell). This is related to the assumption of normality because if both variables are normally distributed than you should have homoscedasticity. There is no formal test for this, but it can be seen graphically. “The bivariate scatterplots between two variables are roughly the same width all over with some bulging toward the middle” (Tabachnick & Fidell, see page 79, 4th ed.).

“Heteroscedasticity is not fatal to an analysis of ungrouped data. The linear relationship between variables is captured by the analysis, but there is even more

predictability if the heteroscedasticity is accounted for. If it is not, the analysis is weakened, but not invalidated” (Tabachnick & Fidell).

****Nonlinearity and heteroscedasticity are assessed simultaneously by viewing the bivariate scatterplots.****

Note: Nonlinearity and heteroscedasticity are corrected by transformations (discussed above). Transforming variables is usually avoided with our research because it distorts interpretability of the findings.

Table for Nonlinearity

What two variables were used in the scatterplot?	Was the scatterplot oval? If yes, stop here.	If the scatterplot was not oval, did you transform the variable(s) to enhance linearity? Yes or No	What transformation did you use?	What were skewness and kurtosis before the transformation?	What were skewness and kurtosis after the transformation?

Date Verified _____

Verifier Name (please print) _____

Table for Heteroscedasticity

What two variables were used in the scatterplot?	Was homoscedasticity present? If yes, stop here.	If heteroscedasticity was present, did you transform the variable(s)?	What transformation did you use?	Describe/draw the pattern of the bivariate scatterplot with the transformed variable(s)?

Date Verified _____

Verifier Name (please print) _____

Step 5: Identify and Deal with Univariate Outliers and Nonnormal variables

A. **Deal with univariate outliers:** After detecting univariate outliers by the methods discussed in Step 2 (using the z-score method and [maybe] Boxplots), you now must deal with them. Going on the assumption that the variable is part of the population you wanted but in the distribution it's an extreme case, which is usually true, you have several options: 1. If you decide that the outlier isn't part of the population, then you can delete it without loss of generalizability of results to your intended population. 2. If you decide that it's part of your population, then you must reduce its impact. You have two options:

i. Change the outliers' scores so that they are still extreme, but less so. Here you could assign the outlying case(s) a raw score on the offending variable that is one unit larger (or smaller) than the next most extreme score in the distribution.

ii. Transform the variable if the outliers seem to part of an overall non-normal distribution, but first check for normality. **I prefer i.**

List all univariate outliers here and how they were handled.

Note: It's wise to check if a univariate outlier is also a multivariate outlier before making any decisions about what to do with it. (More on multivariate outliers below.)

What variable(s)?	Case Number	Z score	Reason for being an outlier (refer to 1-4 in Step 2C a).	How handled (deleted, changed,
-------------------	-------------	---------	--	--------------------------------

				*transformed)

*Only after a check for normality is performed. I prefer leaving variables untransformed because of the issues with interpretability of transformed variables.

Date Verified _____

Verifier Name (please print) _____

B. Check skewness and kurtosis, probability plots:

Normality: The data need to follow a normal distribution in order for most analyses to work properly. Even in situations where normality isn't required if normality exists it makes for a stronger assessment. There are two aspects to normality of a distribution: skewness and kurtosis. Both must be tested before normality can be established.

Note: Test for skewness and kurtosis simultaneously.

- i. **Skewness:** This describes how unevenly the data are distributed with a majority of scores piled up on one side of the distribution and a few stragglers off in one tail of the distribution. Skewness is often, but not always caused by outliers, which hopefully were taken care of in Step 5A.

Skewness test: In SPSS ANALYZE/descriptives/frequencies/ statistics will get you the skewness and the standard error for skewness. Divide the skewness value by the standard error for skewness. You get the z score for skewness. If the number is greater than 3.3 , you have a problem.

Note: Skewness tends to have more influence on analyses than kurtosis.

- ii. **Kurtosis:** This describes how "peaked" or "flat" a distribution is. If too many or all of the scores are piled up on or around the mean then the distribution is too peaked and it's nonnormal, vice versa for when a distribution is too flat.

Kurtosis test: In SPSS ANALYZE/descriptives/frequencies/ statistics will get you the kurtosis and the standard error for kurtosis. Divide the

kurtosis value by the standard error for kurtosis. You get the z score for kurtosis. If the number is greater than 3.3, you have a problem.

Note: The larger the sample size you're using the more likely you'll get violations of skewness and/or kurtosis with small deviations. With larger sample sizes you may want to use a less conservative number than a z -score of 3.3 (pick a higher number).

C. Transform variables (if desired): Data transformation is discussed in Step 4 under "nonlinearity." Because transforming variables hinders interpretability (what's the log of gratitude?), we'll likely not be doing this.

D. Check results of transformations: Same as C above.

Note: Complete the first two columns in the Table for Skewness and Table for Kurtosis even if you don't transform the variables.

Table for Skewness

What variable(s) were significant for skewness?	What was skewness before the transformation?	What transformation did you use?	What was skewness after the transformation?

Date Verified _____

Verifier Name (please print) _____

Table for Kurtosis

What variable(s) were significant for kurtosis?	What was kurtosis before the transformation?	What transformation did you use?	What was kurtosis after the transformation?

Date Verified _____

Verifier Name (please print) _____

Step 6: Identify and Deal with Multivariate Outliers

Multivariate Outliers are found by first computing a Mahalanobis Distance for each case, and once that's done the Mahalanobis scores are **screened in the same manner that univariate outliers are screened.**

To compute Mahalanobis distance in SPSS use ANALYZE/regression/linear. Use the ID variable as the DV and all variables that need to be screened as IVs. Check the Mahalanobis box under SAVE/ Distances. There should be a new variable saved in your data set. For grouped data the Mahalanobis distances must be computed separately for each group. Review the Critical Values of the Chi Square table. The df column refers to the number of predictors. Use $p < .001$ as a conservative probability estimate for a case being an outlier. For example, if you have 10 predictors, the Chi Square critical value is 29.588. If a case has a Mahalanobis Distance *greater* than 29.588, it's a multivariate outlier.

“Once multivariate outliers are identified, you need to discover why these cases are extreme. It's important to identify the variables on which the cases are deviant for three reasons. First, this procedure helps you decide whether the case is properly part of your sample. Second, if you are going to modify scores instead of delete cases, you have to know which scores to modify. Third, it provides an indication of the kinds of cases to which your results do not generalize” (Tabachnick & Fidell).

“If there are only a few multivariate outliers, it is reasonable to examine them individually. If there are several, you can examine them as a group to see if there are any

variables that separate the group of outliers from the rest of the cases” (Tabachnick & Fidell).

“Although the number of possible multivariate outliers is often substantially reduced after transformation or alteration of scores on variables, there are sometimes a few cases that are still far away from the others. These cases are usually deleted. If they are allowed to remain, it is with the knowledge that they may distort the results in almost any direction” (Tabachnick & Fidell).

List all Multivariate outliers here and how they were handled.

What IVs were used in the analysis?	Case Number	Critical Value from Chi Square table	Mahalanobis Score	How handled (deleted, *transformed)

--	--	--	--	--

*Only after a check for normality is performed. I prefer leaving variables untransformed because of the issues with interpretability of transformed variables.

Date Verified _____

Verifier Name (please print) _____

Important: You also need to identify the variables causing the outliers and describe the multivariate outliers.

Identify the variables causing the outliers: Create a dummy variable for cases without multivariate outliers (coded as “0”) and cases with multivariate outliers (coded as “1”).

Did you do this? Yes No

“The dummy variable is then used as a grouping DV in discriminant-function analysis or logistic regression, or as the DV in regression. Within these techniques, **stepwise procedures** are useful for identifying the variables that distinguish outliers from the other cases. Variables on which the outlier(s) differ from the rest of the cases enter stepwise progression of the equation; the remaining variables do not. Once those variables are identified, means on those variables for outlying and nonoutlying cases should be obtained” (Tabachnick & Fidell). Get means in SPSS via ANALYZE/descriptive/descriptives.

What variables distinguish the outliers from the rest of the cases?	What are the means on those variables for the outlying cases?	What are the means on those variables for the nonoutlying cases?

Date Verified _____

Verifier Name (please print) _____

Describe the multivariate outliers:

Date Verified _____

Verifier Name (please print) _____

Step 7: Evaluate Variables for Multicollinearity and Singularity

Multicollinearity and Singularity: Both deal with correlations among variables. If you have a correlation between two variables that is .90 or greater they're multicollinear. If one of the variables (e.g., SWB) is a combination of two or more of the other variables (e.g., PA, NA, and LS), you have singularity, and the variables are redundant. Either way you cannot have variables that are multicollinear or singular in the same analysis because the analysis won't work. There are two things you can do to find out: 1. Run bivariate correlations between all of your variables and ensure that all correlations are $\leq .90$. If so, you're safe. 2. Run your analysis and see if you get an error/warning message telling you that you have multicollinearity/singularity.

Note: The only way to fix multicollinearity or singularity is to drop one of the variables from the analysis.

Date Verified _____

Verifier Name (please print) _____