Data Checks

After you have compiled your data, you have your raw data. You will need to **clean** or **edit** this data to make sure it is as valid as possible before starting to make connections between your findings. Here is a checklist below that can help to guide you through what to check after compiling your initial data.

Ensure you have:
(a) **Entered** your data correctly.
(b) **Transferred** your data correctly.
(c) **Analyzed** your data correctly.

This requires screening your work to check for data entry, measurement, and integration errors in order to identify what is defective about your data.

What, specifically, should I look for?

- misspellings and incorrect formatting
- having too little or too much data
- inconsistent findings that are far beyond the normal distribution of data (outliers)
- patterns that look strange or analysis results that do not look right
- missing data
- errors such as misunderstood answers or typos
- truthful answers that nonetheless are unusual compared to the average answer (outliers)
- valid records
- data with which you may need to make a judgement call regarding how to relate to the data during certain phases of your research process
- "human error" in measurement design and execution.
- use of correct units (for example, meters or feet)
- ranges such as the age of subjects across questionnaires
- inconsistencies in the sources of data collection
- errors in spelling
- the loss of pieces of data

How do I look for these things?

To find these possible data errors:

1. Save a copy of the original, raw data. You will wantto refer to the raw data to better understand sources of any mistakes or anomalies.
2. Skim the data.

Data Checks

3. Pick “random” cases (rows) or variables (columns) to identify errors
    - This is especially useful if combining multiple data sets as a way to ensure proper alignment
4. Run descriptives
    - This is a very **important** step.
    - Running descriptives on the data will not only provide you with a good way to see errors in the entire dataset, but will also help you familiarize yourself with your data.

Some descriptives to run:

(a) ***Means, Range, Mode, Median:*** this will give you a sense of the distribution of your data and help you identify any outliers.

Example:
Let’s say you are calculating the mean score on a depression measure. You find the mean score in your sample to be 5.94. Without looking at other descriptive data, you use this in future calculations.
This depression measure uses a Likert scale (1-7), so the highest possible score on any one question is 7 and the lowest possible score on any one question is 1. Running descriptives, such as the range of data, on each question shows you that on one question, a participant’s score was entered as a “30.” This is clearly a mistake, and this high score will affect other analyses and calculations, such as the mean.

(b) ***Number of participants answering certain questions:***this will help you identify any missing data or participant attrition, especially if you have “branching” questions.

Yes/No Questions

Example:
You ask your 100 participants, “Are you currently employed?” and find that 85 say yes and 15 say no. You can clearly see that you have no missing data for this question (85+15 gives you 100).
On a separate question you ask, “Do you live with someone?” and find that 65 answered “yes” and 12 answered “no.” You can clearly see that you have data missing here from 23 participants (65+12 gives you 77 (out of 100).

Crosstabs

Example:
Let’s say you ask participants, “Are you currently employed?”
Then, depending on their answer, participants are given one of these follow-up questions:

Data Checks

If yes, how long have you been working at your current place of employment?
If no, how long have you been searching for a job?
You run crosstabs on these questions to find how many people answered each question. You see the following:

	Employed?	
	Yes	No
If yes, how long have you been working at your current place of employment?	n=76	n=3,
If no, how long have you been searching for a job?	n=9	n=12

Further examination of this table alerts you to possible mistakes in the data, evident in the cells in red. You see from this that although 15 participants answered "no" to whether they were employed, 3 of these participants answered the follow up question for the "yes" answer. Similarly, although 85 participants answered "yes" to whether they were employed, 9 of these participants answered the follow-up question for the "no" answer.

What do I do after I identify errors in the data?

1. ***Go back to your original, raw data.***
   This may be a file downloaded from a computer survey application, or the actual paper questionnaires that your participants filled out in person.

   This will help you identify the source of the error--whether it was an error in the survey itself (either participant error or error in survey flow or design), or an error in the way you entered or coded your data.

2. ***After identifying errors, correct any mistakes in the dataset*** (remember, always keep a copy of the original in case other issues come up!) by:
   (a) Replacing the incorrect values with the correct values if you are able to find them
   (b) Replace the incorrect values by removing them from the dataset and treating them as "missing values." Refer to the missing values handout for strategies on how to handle missing data.