Using SPSS to Restructure Data from Computerized Experiments

Jamie DeCoster
Department of Psychology
University of Alabama
348 Gordon Palmer Hall
Box 870348
Tuscaloosa, AL 35487-0348

December 29, 2005

If you wish to cite the contents of this document, the APA reference for them would be

DeCoster, J. (2004). Using SPSS to Restructure Data from Computerized Experiments. Retrieved (month, day, and year you downloaded this file, without parentheses) from http://www.stat-help.com/notes.html

Table of Contents

Data From Computerized Experiments

There are a host of benefits of using a computer to conduct your study. They let you easily present complex stimuli, such as pictures, sounds, and video. They allow you to standardize the presentation of your study across participants. They make your study portable, allowing you to conduct the same study on multiple computers using a single experimenter. They allow you to randomize the presentation of stimuli, allowing you to avoid counterbalancing factors in your design. They will even automatically record your data, often with greater precision and accuracy than could be obtained without the computer.

One problem with computerized experiments, however, is that they often record data in a way that makes it difficult to analyze using standard statistical software. Specifically, computerized experiments will commonly record data in *univariate format*, where each line in the data set corresponds to a different trial. Data from a single subject will therefore be spread over a number of different rows. Figure 1 below is an example of a data set represented in a univariate format.

Subject	Cond	Stim	RT
1	A	CAT	453
1	A	DOG	594
1	A	ELK	256
1	A	FOX	610
2	B	CAT	405
2	B	DOG	773
2	B	ELK	594
2	B	FOX	663
3	B	CAT	835
3	B	DOG	539
3	B	ELK	745
3	B	FOX	404
4	B	CAT	963
4	B	DOG	638
4	B	ELK	944
4	B	FOX	734
5	A	CAT	656
5	A	DOG	503
5	A	ELK	698
5	A	FOX	468

Figure 1. A data set in univariate format.

We can that the study from which this data was collected had one between-subjects factor (condition) and one within-subjects factor (stimulus). The easy way to distinguish these is to notice that the between-subjects factor has the same value on every trial for a given participant, while the value of the within-subjects factor varies from trial to trial. Computer programs typically write files in this format because each line can be put into the file as soon as the trial is

completed. This means that the computer only ever has to keep four values in mind – one for each of the variables in the data set.

Data analytic procedures will commonly call for a data set to be stored in *multivariate format*, where each line in the data set corresponds to a different subject. Responses on the different trials are stored as different variables. Figure two below is an example of a data set in multivariate format.

Subject	Cond	CatRt	DogRt	ElkRT	FoxRT
1	A	453	594	256	610
2	B	405	773	594	663
3	B	835	539	745	404
4	B	963	638	944	734
5	A	656	503	698	468

Figure 2. A data set in multivariate format.

Notice that the data sets in Figure 1 and Figure 2 are both representing exactly the same thing, even though they do it with entirely different cases and variables. In both cases, you can see that the RT of subject 2 to "Dog" is 773, the RT of subject 3 to "Cat" is 835, and the RT of subject 5 to "Fox" is 468. You can similarly see that subjects 1 and 5 are in condition A, while subjects 2, 3, and 4 are in condition B. However, the two data sets are not equally useful. It is much easier to analyze the above data when it is in multivariate format than when it is in univariate format. Both repeated measures ANOVA and MANOVA procedures available in SPSS expect your data to have a multivariate format. The primary goal of these notes is therefore to explain how you can turn a data set like that presented in Figure 1 into one like that presented in Figure 2.

Step 1: Determine the Trial Sets You Want to Combine

The very first thing that you need to do is decide what trials you want to include in your data set. Sometimes a single experiment will collect several different types of responses. For example, during an experiment you might first show people a list of words, then ask them to recall as many words as they can from the list, and finally have them perform a multiple-choice recognition task for items on the list. Restructuring this data set will be easier if you restructure the data from the recall task and the data from the recall task separately. This does not mean that you can't include all the information in a single data set - at the very end you can merge all the data sets that you create. It is simply easier to keep them separate while you are restructuring them. A good guideline is to separate trials that record either different types of information or that record information in different sets of variables. Usually you will not need to separate different blocks of trials from the same task. Steps 2 through 4 below should be performed on each set of trials.

Step 2: Prepare the Data File in Excel
Before you try to restructure your data set there are several edits you should make to it in Excel. First, you should decide what variables you want to include in your data set. Experimental programs such as Eprime will often record information you won't need for your analysis, such as

the time of day a subject was run, or the number used as the random seed. Converting a data set from univariate to multivariate format will naturally increase the number of variables in your data set, so it's best to delete those you don't need at the start so that you don't exacerbate the problem. Deleting a variable in Excel is very easy. All you need to do is click on the column heading (which selects the column) and then choose *Edit → Delete*.

Second, you should check the names of the variables that you want to keep to make sure that they are compatible with SPSS. The names of your variables should be placed in the first row of the data set. Below are some rules SPSS has about variable names.

- The name cannot have more than 8 characters
- The first character must be a letter
- The second through eighth characters can be letters, numbers, a period, an underscore, or the symbols $, #, or @
- The name cannot include spaces

In addition to these rules, we would also suggest that the name of any within-subjects variables be kept to four characters or less. SPSS will add characters to the names of within-subject variables after you restructure your data, so you need to leave enough space.

Third, you should delete any trials that are not in the set you are working with. This is slightly more difficult than deleting unnecessary variables, since you will likely need to delete several lines from each subject. The easy way to do this is to first sort the rows in your data set by the trial number. This will put all of the trials that you want to work with in a single place in the file. That way you can delete the ones you don't want all at once, instead of having to delete them for each subject individually. To delete the trials you don't want you should first select them, and then choose *Edit → Delete*. If your data set doesn't already have a trial number variable you can add one to the data set using the Excel formula

= mod(row( )-2, *number*)

where you replace *number* with the number of trials you have for each subject.

You should make sure that the trials you decide to keep include some sort of subject number. Each participant in your study should have a unique subject number – two different people shouldn't have the same number. If you want to combine different data sets from the same study, each data set will need to include this subject number so that SPSS will know which data came from which subject.

Step 3: Load the Univariate File in SPSS

Once you have prepared the file in Excel you should save it as an Excel version 4.0 worksheet. To do this you should choose *File → Save as*, click the pull-down menu next to the *Save as type* box, and the choose *Microsoft Excel 4.0 Worksheet*. It is generally easier to export Excel version 4.0 worksheets to other programs, so we typically choose to save data sets in this format when our goal is to load it into SPSS.

After saving the Excel file you must then load it into SPSS. To do this you take the following steps.

- Choose *File* → *Open* → *Data*
- Click the pull-down menu next to the *Files of type* box and select *Excel (*.xls)*
- Locate the Excel file you wish to load and enter its name in the *File name* box.
- Click the *Open* button.
- Check the box next to *Read variable names*
- Click the *OK* button

After you load the data set into SPSS you should save it in SPSS format. To do this you just have to choose *File* → *Save.*

Step 4: Sort the Data Set

You should next sort the cases in the data set so that they are first ordered by the subject number, and then ordered by the levels of the within-subjects variables. To do this in SPSS you would take the following steps.

- Choose *Data* → *Sort Cases*
- Move the variable corresponding to the subject number to the *Sort by:* box.
- Move the variables corresponding to the within-subjects variables to the *Sort by:* box. Be sure to note the order in which you add these variables to the box because this will determine the order of the conditions in the multivariate data set.
- Click the *OK* button.

Step 5: Restructure the Data Set

This is the most important step of the process, where we actually combine the information contained in multiple trials in the univariate data set to form the multivariate data set. This section will assume that the design of the study is the same for each subject, although it is perfectly acceptable if the exact order in which the trials were run was randomized. To restructure the data set you take the following steps.

- Load the univariate data set into SPSS.
- Choose *Data* → *Restructure*.
- Click the radio button next to *Restructure selected cases into variables.*
- Click the *Next* button.
- Move the variable containing the subject number to the *Identifier Variable(s)* box.
- Click the *Next* button.
- Click the *Finish* button.

Once you do this, SPSS will combine all of the data that comes from each subject into a single case in the data set. Any variables that always have the same value in all of the rows from a

single subject (i.e., between-subject variables) will be represented by a single variable in the new data set. Any variables whose values change from row to row (i.e., within-subject variables) will be represented by a set of variables in the new data set. This set will contain a number of variables equal to the number of trials. Each of these variables will start with the name of the original variable and end with an indicator of which trial the value was taken from. For example, if we had SPSS restructure the data set provided in Figure 1, we would end up with the data set illustrated below in Figure 3.

Subject	Condition	Stim.1	Stim.2	Stim.3	Stim.4
1	A	CAT	DOG	ELK	FOX
2	B	CAT	DOG	ELK	FOX
3	B	CAT	DOG	ELK	FOX
4	B	CAT	DOG	ELK	FOX
5	A	CAT	DOG	ELK	FOX

RT.1	RT.2	RT.3	RT.4
453	594	256	610
405	773	594	663
835	539	745	404
963	638	944	734
656	503	698	468

Figure 3. The data in Figure 1 after being restructured by SPSS.

Since we had already sorted the trials in the initial data set by the name of the stimulus, you can see that the particular trial numbers each correspond to a specific condition of our study. This means that we actually don't need the information provided by the variables **Stim.1**, **Stim.2**, **Stim.3**, and **Stim.4** since it will always be the same for every subject. In this case you might choose to delete these variables from the data set and instead rename the variables **RT.1**, **RT.2**, **RT.3**, and **RT.4** so that they indicate the condition of the within-subjects factor to which they refer. Doing this will provide you with the restructured data set illustrated in Figure 2 above.

If you have replications of your within-subjects conditions then you will have several different variables that will correspond to a single condition. They will all be located together in the data set, making it easy to calculate summary statistics (such as means) to use in your analyses. If you have multiple within-subject factors then you will have a variable in your data set corresponding to each specific combination of these factors. For example, if we had chosen to cross the stimulus type factor used in the examples above (with 4 levels) with a factor indicating whether the stimulus was presented in upper case or lower case (2 levels), we would expect our resulting data set to have $4 \times 2 = 8$ different reaction times. Which variables correspond to which combinations would depend on the order you used when sorting the data. The order of the variables will be exactly the same as the order of the rows after sorting.

Instead of having SPSS identify the trial by a specific number, you can also have it automatically label the trial based on the entry in a specific variable. For example, we might want to have

SPSS automatically label the RT variables based on the stimulus that was used. To do this you would take the following steps.

- Load the univariate data set into SPSS.
- Choose *Data* → *Restructure*.
- Click the radio button next to *Restructure selected cases into variables.*
- Click the *Next* button.
- Move the variable containing the subject number to the *Identifier Variable(s)* box.
- Move the variable containing the labels to the *Index Variable(s)* box.
- Click the *Next* button.
- Click the *Finish* button.

If you only have a single response variable, the names of the variables in the transformed SPSS data set will be set equal to the levels of the index variable. For example, the data set illustrated only has a single response variable (**RT**), so after restructuring the data set using **Stim** as an index variable we would get the data set illustrated in Figure 4 below.

Subject	Condition	Indcat	Inddog	Indelk	Indfox
1	A	1	1	1	1
2	B	1	1	1	1
3	B	1	1	1	1
4	B	1	1	1	1
5	A	1	1	1	1

Cat	Dog	Elk	Fox
453	594	256	610
405	773	594	663
835	539	745	404
963	638	944	734
656	503	698	468

Figure 4. Example of restructuring a data set using an index variable.

You can see that the variables that were labeled **RT.1**, **RT.2**, **RT.3**, and **RT.4** above would now be named **CAT**, **DOG**, **ELK**, and **FOX**. The data set would also contain an "indicator" for each of our variables. Each indicator has a value of 1 if that particular subject has a value for the corresponding response variable, and a value of 0 if it does not.

If you have multiple response variables, on the other hand, SPSS will tack the label onto the end of the variable name. For example, if our example data set also included the accuracy on each trial in a variable named **ACC**, the resulting data set would contain the variables **RT.CAT**, **RT.DOG**, **RT.ELK**, **RT.FOX**, **ACC.CAT**, **ACC.DOG**, **ACC.ELK**, and **ACC.FOX**.

Note that if a variable name generated by SPSS would ever have more than 8 characters, SPSS will not use the name and instead will replace it with a generic name like **VAR00001.** It is

therefore a good idea to keep the names of your response variables and the values of any index variables as short as possible so that they can be combined to produce useful names. If you cannot you can still identify which variables are which because SPSS will assign the name that it wanted to use as the label for the variable.

Step 6: Merge the Multivariate Data Sets

If you chose to split your file to make restructuring your data sets easier, the final step you must take is merge those files into a complete data set. You can take the following steps to merge two files in SPSS.

- Load one of the data sets into SPSS.
- Choose *Data* → *Merge files* → *Add variables.*
- Enter the name of a second data set in the *File name* box.
- Click the *OK* button.
- Check the box next to *Match cases on key variables in sorted files.*
- Move the variable containing the subject number to the *Key Variables* box.
- Click the *OK* button.

The above procedure can be repeated multiple times to merge the data contained in multiple files.