

Data Analysis in SPSS

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

Heather M. Claypool
Department of Psychology
Miami University of Ohio
136 Benton Hall
Oxford, OH 45056

February 21, 2004

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

DeCoster, J., & Claypool, H. M. (2004). Data Analysis in SPSS. Retrieved <month, day, and year you downloaded this file> from <http://www.stat-help.com/notes.html>

All rights to this document are reserved

Table of Contents

Introduction	1
Interactive Mode versus Syntax Mode	2
Descriptive Statistics	4
Transformations	5
Compute	5
Recode	6
Reverse coding	7
Selecting Cases	9
t Tests	11
One-sample t-test	11
Independent-samples t-test	11
Paired-samples t-test	12
Analysis of Variance (ANOVA)	13
One-way between-subjects ANOVA	13
Multifactor between-subjects ANOVA	14
One-way within-subjects ANOVA	17
Multifactor within-subjects ANOVA	19
Mixed ANOVA	20
MANOVA	23
Performing a MANOVA using interactive mode	24
Performing a MANOVA using syntax	25
Interpreting MANOVA output	26
Correlation	28
Pearson correlation	28
Point-biserial correlation	28
Spearman rank correlation	29
Regression	30
Simple Linear Regression	30
Multiple Regression	32
Multiple regression with interactions	33
Polynomial regression	36
Simultaneously testing categorical and continuous IVs	37
Mediation	39
Chi-Square Test of Independence	40
Logistic Regression	42
Reliability	45
Calculating reliability from parallel measurements	45
Calculating reliability from internal consistency	46
Inter-rater reliability	47
Factor Analysis	50
Vectors and Loops	54
Vectors	54
Loops	55
The Power of Combining Loops and Vectors	56

INTRODUCTION

These notes are designed to provide readers with a practical overview of how to perform data analysis in SPSS. These notes are divided into sections, each of which discusses what a particular SPSS procedure is used for, what specific steps you need to take to perform the analysis, and how you should interpret the resulting output. We also discuss any “tricks” that we have come across in our own use of these procedures to deal with common difficulties.

The notes are written with the assumption that the reader has a basic knowledge of statistics, such as might be expected from a typical graduate student in the social sciences. Beyond that we do not assume that the reader has any specific familiarity with the specific types of analyses we discuss. Our explanations and descriptions were written to be thorough and complete so that, for example, someone unfamiliar with logistic regression would be able to read through that section and understand not only what the procedure is used for, but would be able to perform a basic logistic regression themselves and understand how to interpret the results.

We would like to thank Diane Mackie for the original commission of these notes for a workshop at the University of California, Santa Barbara. We would also like to thank the original participants (Melissa Ryan, Angela Maitner, Wesley Moons, and Sara Crump) for their excellent questions and comments during our presentation. We both came away from the workshop knowing more than we did going into it.

INTERACTIVE MODE VERSUS SYNTAX MODE

There are two basic ways that you can work with SPSS. Most users typically open up an SPSS data file in the data editor, and then select items from the menus to manipulate the data or to perform statistical analyses. This is referred to as *interactive mode*, because your relationship with the program is very much like a personal interaction, with the program providing a response each time you make a selection. If you request a transformation, the data set is immediately updated. If you select an analysis, the results immediately appear in the output window.

It is also possible to work with SPSS in *syntax mode*, where the user types code in a syntax window. Once the full program is written, it is then submitted to SPSS to get the results. Working with syntax is more difficult than working with the menus, because you must learn how to write the programming code to produce the data transformations and analyses that you want. However, certain procedures and analyses are only available through the use of syntax. For example, vectors and loops (described later) cannot be used in interactive mode. You can also save the programs you write in syntax. This can be very useful if you expect to perform the same or similar analyses multiple times, since you can just reload your old program and run it on your new data (or your old data if you want to recheck your old analyses). If you would like more general information about writing SPSS syntax, you should examine the *SPSS Base Syntax Reference Guide*.

Whether you should work in interactive or syntax mode depends on several things. Interactive mode is easier and generally quicker if you only need to perform a few simple transformations or analyses on your data. You should therefore probably work interactively unless you have a specific reason to use syntax. Some reasons to choose syntax would be:

- You need to use options or procedures that are not available using interactive mode.
- You expect that you will perform the same procedures on several different data sets and want to save a copy of the program code so that it can easily be re-run.
- You need to perform a large number of similar transformations, such that using vectors and loops would benefit you.
- You are performing a very complicated set of procedures, such that it would be useful to document all of the steps leading to your results.

Whenever you make selections in interactive mode, SPSS actually writes down syntax code reflecting the menu choices you made in a "journal file." The name of this file can be found (or changed) by choosing **Edit → Options** and then selecting the **General** tab. If you ever want to see or use the code in the journal file, you can edit the journal file in a syntax window.

SPSS also provides an easy way to see the code corresponding to a particular menu function. Most selections include a **Paste** button that will open up a syntax window containing the code for the function, including the details required for any specific options that you have chosen in the menus.

Finally, you can have SPSS include the corresponding syntax in the output whenever it runs a statistical analysis. To enable this

- Choose **Edit → Options**.
- Select the **Viewer** tab.
- Check the box next to **Display commands in the log**.
- Click the **OK** button.

DESCRIPTIVE STATISTICS

Analyses often begin by examining basic descriptive-level information about data. The most common and useful descriptive statistics are

- Mean
- Median
- Mode
- Frequency
- Quartiles
- Sum
- Variance
- Standard deviation
- Minimum/Maximum
- Range

Note: All of these are appropriate for continuous variables, and frequency and mode are also appropriate for categorical variables.

If you just want to obtain the mean and standard deviation for a set of variables

- Choose **Analyze → Descriptive Statistics → Descriptives**.
- Move the variables of interest to the **Variable(s)** box.
- Click the **OK** button.

If you want to obtain any other statistics

- Choose **Analyze → Descriptive Statistics → Frequencies**.
- Move the variables of interest to the **Variable(s)** box
- Click the **Statistics** button.
- Check the boxes next to the statistics you want.
- Click the **Continue** button.
- Click the **OK** button.

TRANSFORMATIONS

Compute

The **Compute** procedure allows the analyst to perform mathematical operations on variables. This can be useful for several reasons. You might want to transform a variable to make its distribution more normal, or to make its relationship to another variable more linear. You also might want to create a variable that represents the average of other variables in the data set. You should use **Compute** whenever you want to assign the value of one variable to be a mathematical function of other variables in the data set.

To perform this type of a transformation

- Choose **Transform → Compute**.
- Type the name of the new variable in the **target variable** box. If you type in the name of a variable that already exists, the transformed values will replace the existing values.
- Type the formula representing what the value of the target variable should be in the **numeric expression** box.
- Click the **OK** button.

The screen that opens up when you choose the **Compute** procedure has three different parts. On the left side of the screen there is a list of the variables in your data set. If you double-click on a variable it will be added to the computation formula listed in the **numeric expression** box. You do not have to use this if you do not want to – you can always just type in the names of the variables in your formula by hand.

In the center of the screen is what appears to be a calculator keypad. This allows you to add specific arithmetic functions to your formula. The keypad contains the following buttons for mathematical computations.

- + (addition)
- - (subtraction)
- * (multiplication)
- / (division)
- ** (raising something to a power)

The calculator keypad also contains the followings functions that allow you to compare variables and values. These are primarily used when you click the **if** button to limit the cases affected by your transformation.

- < (less than)
- > (greater than)
- <= (less than or equal to)
- >= (greater than or equal to)
- = (equal to)
- ~= (not equal to)
- & (Boolean AND)

- | (Boolean OR)
- ~ (Boolean NOT)

Finally, the keypad contains a button with parentheses (). Clicking this button inserts a pair of parentheses into your formula, which can be used to tell SPSS which parts of your formula should be evaluated first.

As with the list of variable names, you do not have to use the keypad to use these functions – you can get the same results by typing the formula you want directly into the box labeled **Numeric expression**.

The right-hand part of the screen contains a list of functions that you can include in your formulas. There is a complete list of the functions that you can use in the box labeled **Functions**. Double clicking a function in this list will add it to your formula, with question marks in place of where you would need to insert the names of actual variable names. This can be handy if you find it difficult to remember the names of the functions you want to use. However, you do not have to use this list if you do not want to – you can type in the functions directly into your formula by hand. Below are some commonly used functions you can use in your formula.

- **mean**(*var1*, *var2*, ...) – provides the average of the variables in the parentheses
- **sum**(*var1*, *var2*, ...) – provides the sum of the variables in the parentheses
- **sqrt**(*var*) – provides the square root of the variable in parentheses
- **lg10**(*var*) – provides the base 10 logarithm of the variable in parentheses
- **ln**(*var*) – provides the natural logarithm of the variable in parentheses

You can right-click any function in the list to find out more about it.

Recode

The **Recode** procedure is typically used with transformations involving categorical variables. It is the best option when you want to create a categorical distinction based on an existing numeric variable (such as a median split), when you want to combine some of the categories in an existing categorical variable, or when you simply want to change the values assigned to an existing categorical variable.

In general it is recommended that you use numbers to code different levels of your categorical variables in SPSS. Although SPSS variables can have letters as values (these types of variables are called “string” variables), there are several important analysis that can only work with numeric variables, even if the variable itself is categorical.

Whenever you work with the **Recode** procedure you must choose whether to recode **Into Same Variables** or to recode **Into Different Variables**. If you choose to recode **Into Same Variables** then the result of your transformation will replace the values of your original variable. If you choose to recode **Into Different Variables** then the result of your transformation will be placed into a new variable that you will have to name. We suggest that you always use **Into Different Variables** so that you can later choose to redo the transformation if you want.

To perform a median split

- Obtain the median of the variable as described in the section on *Descriptive Statistics*.
- Choose **Transform → Recode → Into different variables**.
- Select the original continuous variable and click the arrow button.
- Type the name of the new categorical variable in the box labeled **Name**.
- Click the **Change** button.
- Click the **Old and New Values** button.
- Click the radio button next to **Range: Lowest through**.
- Type the value of the median in the open box on the left side.
- Type the value for the low category in the box labeled **Value** on the right side.
- Click the **Add** button
- Click the radio button next to **All other values**.
- Type the value for the high category in the box labeled **Value** on the right side.
- Click the **Add** button.
- Click the **Continue** button.
- Click the **OK** button.

To change the values of an existing categorical variable

- Choose **Transform → Recode → Into different variables**.
- Select the original categorical variable and click the arrow button.
- Type the name of the new categorical variable in the box labeled **Name**.
- Click the **Change** button.
- Click the **Old and New Values** button.
- You must tell SPSS how the values of the old categorical variable map onto the values of the new variable. For each level of the old variable
 - Make sure the radio button next to **Value** is selected.
 - Type the value of the original variable in the open box on the left side.
 - Type the corresponding value of the new variable in the box labeled **Value** on the right side. It's perfectly acceptable to assign the same new value to several different old values.
 - Click the **Add** button.
- Click the **Continue** button.
- Click the **OK** button.

One nice feature of the **Recode** function is that it remembers the last set of recodes that you asked it to perform. If you recode a variable and then select the recode function again, you'll see the **Old → New** combinations from the prior transformation. This can be very useful if you need to apply the same recoding to a number of different variables. In this case you could simply change the input variable and output variable and then run the transformation, leaving the old and new values the same.

Reverse coding

Reverse coding is a procedure where some questions in a survey are worded such that high values of a theoretical construct is reflected by high scores on the item, while other questions are worded such that high values of the same construct is reflected by low scores on the item.

Researchers do this to encourage respondents to actually pay attention to the questions they are reading. Unfortunately this means that you can't determine the overall score for the scale simply by averaging the items. Instead you must first transform the items so that they are all oriented in the same direction. For example, all items might be scored such that large values indicate more of the construct. To do this, you would want to reverse code the items where small values indicated a greater amount of the construct. So, if the questions in the scale had values of 1 to 7, you would reverse code an item by changing its values in the following way:

Old Value	New value
1	7
2	6
3	5
4	4
5	3
6	2
7	1

While it would be possible to perform this transformation using the **Recode** procedure, there is a simple formula you can use to do the same thing using the **Compute** procedure. The formula is

$$\text{new value} = (\text{scale minimum} + \text{scale maximum}) - \text{old value}$$

In the current example, the scale minimum is 1 and the scale maximum is 7. Therefore, the formula we'd need to use would be **8 – old value**. You can verify for yourself that this will produce the transformation described above. The formula will work for any possible scale minima and maxima, even if the scale has values less than zero.

SELECTING CASES

Sometimes a researcher may wish to restrict an analysis to only a subset of the data file. For example, a researcher might want to do an analysis on only the female participants, only on American citizens, or only on those who disagreed with a presented message. You can have SPSS only perform analyses on a subset of cases meeting specified criteria by selecting cases.

For example, suppose you believed that an experimental manipulation would work better on female participants. You might therefore want to do an analysis only on the female participants. Assuming that you had a variable named **gender** in your data set where men had a value of 1 and women had a value of 2, you could select only women for analysis in SPSS by taking the following steps

- Choose **Data**→ **Select Cases**.
- Click the radio button next to **If condition is satisfied**.
- Click the **If** button.
- Type **gender=2** in the popup window.
- Click the **Continue** button.
- Click the **OK** button.

After running this procedure, the dataset will have slashes crossing out the cases that are excluded (in this case, it should cross out all the male participants). Any further analyses on this data set (until you issue another **Select Cases** command or load another data set) will be performed solely on the female participants.

An analyst can restrict the data set in more complicated ways by using the Boolean operator AND (represented by the symbol **&**) or the Boolean operator OR (represented by the symbol **|**). The **&** symbol tells SPSS that a case has to meet two specific criteria to be included in the analysis, while the **|** symbol tells SPSS that it should include a case if it meets either of two criteria. As an example, let us assume that the data set included a variable named **class** where 1 = freshmen, 2 = sophomore, 3 = junior, and 4 = senior. If we only wanted to include cases that represented female juniors we would perform the following steps.

- Choose **Data**→ **Select Cases**.
- Click the radio button next to **If condition is satisfied**.
- Click the **If** button.
- Type **(gender=2) & (class=3)** in the popup window.
- Click the **Continue** button.
- Click the **OK** button.

If we wanted to include cases that were either female or juniors (so that the analysis would include all females and any males who happened to be juniors) we would take the following steps.

- Choose **Data**→ **Select Cases**.
- Click the radio button next to **If condition is satisfied**.

- Click the **If** button.
- Type **(gender=2) | (class=3)** in the popup window.
- Click the **Continue** button.
- Click the **OK** button.

You may notice that one of the buttons on the selection keypad looks like \neq . This symbol stands for “not equal to.” Sometimes it is easier to identify what cases to exclude than it is to identify what cases to include. For example, you could select everyone in the data set **except** sophomores by taking the following steps.

- Choose **Data** → **Select Cases**.
- Click the radio button next to **If condition is satisfied**.
- Click the **If** button.
- Type **class \neq 2** in the popup window.
- Click the **Continue** button.
- Click the **OK** button.

T TESTS

Many analyses in psychological research involve testing hypotheses about means or mean differences. Below we describe the SPSS procedures that allow you to determine if a given mean is equal to either a fixed value or some other mean.

One-sample t-test

You perform a one-sample t-test when you want to determine if the mean value of a target variable is different from a hypothesized value.

To perform a one-sample t-test in SPSS

- Choose **Analyze** → **Compare Means** → **One-sample t-test**.
- Move the variable of interest to the **Test variable(s)** box.
- Change the **test value** to the hypothesized value.
- Click the **OK** button.

The output from this analysis will contain the following sections.

- **One-Sample Statistics**. Provides the sample size, mean, standard deviation, and standard error of the mean for the target variable.
- **One-Sample Test**. Provides the results of a t-test comparing the mean of the target variable to the hypothesized value. A significant test statistic indicates that the sample mean differs from the hypothesized value. This section also contains the upper and lower bounds for a 95% confidence interval around the sample mean.

Independent-samples t-test

You perform an independent-samples t-test (also called a between-subjects t-test) when you want to determine if the mean value on a given target variable for one group differs from the mean value on the target variable for a different group. This test is only valid if the two groups have entirely different members. To perform this test in SPSS you must have a variable representing group membership, such that different values on the group variable correspond to different groups.

To perform an independent-samples t-test in SPSS

- Choose **Analyze** → **Compare Means** → **Independent-sample t-test**.
- Move the target variable to the **Test variable(s)** box.
- Move the group variable to the **Grouping variable** box.
- Click the **Define groups** button.
- Enter the values corresponding to your two groups you want to compare in the boxes labeled **group 1** and **group 2**.
- Click the **Continue** button.
- Click the **OK** button.

The output from this analysis will contain the following sections.

- **Group Statistics**. Provides descriptive information about your two groups, including the sample size, mean, standard deviation, and the standard error of the mean.

- **Independent Samples Test.** Provides the results of two t-tests comparing the means of your two groups. The first row reports the results of a test assuming that the two variances are equal, while the second row reports the results of a test that does not assume the two variances are equal. The columns labeled **Levene's Test for Equality of Variances** report an F test comparing the variances of your two groups. If the F test is significant then you should use the test in the second row. If it is not significant then you should use the test in the first row. A significant t-test indicates that the two groups have different means. The last two columns provide the upper and lower bounds for a 95% confidence interval around the difference between your two groups.

Paired-samples t-test

You perform a paired samples t-test (also called a within-subjects t-test) when you want to determine whether a single group of participants differs on two measured variables. Probably the most common use of this test would be to compare participants' response on a measure before a manipulation to their response after a manipulation. This test works by first computing a difference score for each participant between the within-subject conditions (e.g. post-test – pre-test). The mean of these difference scores is then compared to zero. This is the same thing as determining whether there is a significant difference between the means of the two variables.

To perform a paired-samples t-test in SPSS

- Choose **Analyze** → **Compare Means** → **Paired-samples t-test**.
- Click the two variables you want to compare in the box on the left-hand side.
- Click the arrow button.
- Click the OK button.

The output from this analysis will contain the following sections.

- **Paired Samples Statistics.** Provides descriptive information about the two variables, including the sample size, mean, standard deviation, and the standard error of the mean.
- **Paired Samples Correlations.** Provides the correlation between the two variables.
- **Paired Samples Test.** Provides the results of a t-test comparing the means of the two variables. A significant t-test indicates that there is a difference between the two variables. It also contains the upper and lower bounds of a 95% confidence interval around the difference between the two means.

ANALYSIS OF VARIANCE (ANOVA)

One-way between-subjects ANOVA

A one-way between-subjects ANOVA allows you to determine if there is a relationship between a categorical *independent variable* (IV) and a continuous *dependent variable* (DV), where each subject is only in one level of the IV. To determine whether there is a relationship between the IV and the DV, a one-way between-subjects ANOVA tests whether the means of all of the groups are the same. If there are any differences among the means, we know that the value of the DV depends on the value of the IV. The IV in an ANOVA is referred to as a *factor*, and the different groups composing the IV are referred to as the *levels* of the factor. A one-way ANOVA is also sometimes called a single factor ANOVA.

A one-way ANOVA with two groups is analogous to an independent-samples t-test. The p-values of the two tests will be the same, and the F statistic from the ANOVA will be equal to the square of the t statistic from the t-test.

To perform a one-way between-subjects ANOVA in SPSS

- Choose **Analyze** → **General Linear Model** → **Univariate**.
- Move the DV to the **Dependent Variable** box.
- Move the IV to the **Fixed Factor(s)** box.
- Click the **OK** button.

The output from this analysis will contain the following sections.

- **Between-Subjects Factors.** Lists how many subjects are in each level of your factor.
- **Tests of Between-Subjects Effects.** The row next to the name of your factor reports a test of whether there is a significant relationship between your IV and the DV. A significant F statistic means that at least two group means are different from each other, indicating the presence of a relationship.

You can ask SPSS to provide you with the means within each level of your between-subjects factor by clicking the **Options** button in the variable selection window and moving your within-subjects variable to the **Display Means For** box. This will add a section to your output titled **Estimated Marginal Means** containing a table with a row for each level of your factor. The values within each row provide the mean, standard error of the mean, and the boundaries for a 95% confidence interval around the mean for observations within that cell.

Post-hoc analyses for one-way between-subjects ANOVA. A significant F statistic tells you that at least two of your means are different from each other, but does not tell you where the differences may lie. Researchers commonly perform post-hoc analyses following a significant ANOVA to help them understand the nature of the relationship between the IV and the DV. The most commonly reported post-hoc tests are (in order from most to least liberal): LSD (Least Significant Difference test), SNK (Student-Newman-Keuls), Tukey, and Bonferroni. The more liberal a test is, the more likely it will find a significant difference between your means, but the more likely it is that this difference is actually just due to chance.

Although it is the most liberal, simulations have demonstrated that using LSD post-hoc analyses will not substantially increase your experimentwide error rate as long as you only perform the post-hoc analyses after you have already obtained a significant F statistic from an ANOVA. We therefore recommend this method since it is most likely to detect any differences among your groups.

To perform post-hoc analyses in SPSS

- Repeat the steps necessary for a one-way ANOVA, but do not press the **OK** button at the end.
- Click the **Post-Hoc** button.
- Move the IV to the **Post-Hoc Tests for** box.
- Check the boxes next to the post-hoc tests you want to perform.
- Click the **Continue** button.
- Click the **OK** button.

Requesting a post-hoc test will add one or both of the following sections to your ANOVA output.

- **Multiple Comparisons.** This section is produced by LSD, Tukey, and Bonferroni tests. It reports the difference between every possible pair of factor levels and tests whether each is significant. It also includes the boundaries for a 95% confidence interval around the size of each difference.
- **Homogenous Subsets.** This section is produced by SNK and Tukey tests. It reports a number of different subsets of your different factor levels. The mean values for the factor levels within each subset are not significantly different from each other. This means that there is a significant difference between the mean of two factor levels only if they do not appear in any of the same subsets.

Multifactor between-subjects ANOVA

Sometimes you want to examine more than one factor in the same experiment. Although you could analyze the effect of each factor separately, testing them together in the same analysis allows you to look at two additional things. First, it lets you determine the independent influence of each of the factors on the DV, controlling for the other IVs in the model. The test of each IV in a multifactor ANOVA is based solely on the part of the DV that it can predict that is not predicted by any of the other IVs.

Second, including multiple IVs in the same model allows you to test for interactions among your factors. The presence of an interaction between two variables means that the effect of the first IV on the DV depends on the level of the second IV. An interaction between three variables means that the nature of the two-way interaction between the first two variables depends on the level of a third variable. It is possible to have an interaction between any number of variables. However, researchers rarely examine interactions containing more than three variables because they are difficult to interpret and require large sample sizes to detect.

Note that to obtain a valid test of a given interaction effect your model must also include all *lower-order* main effects and interactions. This means that the model has to include terms representing all of the main effects of the IVs involved in the interaction, as well as all the

possible interactions between those IVs. So, if you want to test a 3-way interaction between variables A, B, and C, the model must include the main effects for those variables, as well as the AxB, AxC, and the BxC interactions.

To perform a multifactor ANOVA in SPSS

- Choose **Analyze** → **General Linear Model** → **Univariate**.
- Move the DV to the **Dependent Variable** box.
- Move all of your IVs to the **Fixed Factor(s)** box.
- By default SPSS will include all possible interactions between your categorical IVs. If this is not the model you want then you will need to define it by hand by taking the following steps.
 - Click the **Model** button.
 - Click the radio button next to **Custom**.
 - Add all of your main effects to the model by clicking all of the IVs in the box labeled **Factors and covariates**, setting the pull-down menu to **Main effects**, and clicking the arrow button.
 - Add each of the interaction terms to your model. You can do this one at a time by selecting the variables included in the interaction in the box labeled **Factors and covariates**, setting the pull-down menu to **Interaction**, and clicking the arrow button for each of your interactions. You can also use the setting on the pull-down menu to tell SPSS to add all possible 2-way, 3-way, 4-way, or 5-way interactions that can be made between the selected variables to your model.
 - Click the **Continue** button.
- Click the **Options** button and move each independent variable and all interaction terms to the **Display means for** box.
- Click the **Continue** button.
- Click the **OK** button.

The output of this analysis will contain the following sections.

- **Between-Subjects Factors.** Lists how many subjects are in each level of each of your factors.
- **Tests of Between-Subjects Effects.** The row next to the name of each factor or interaction reports a test of whether there is a significant relationship between that effect and the DV, independent of the other effects in the model.

You can ask SPSS to provide you with the means within the levels of your main effects or your interactions by clicking the **Options** button in the variable selection window and moving the appropriate term to the **Display Means For** box. This will add a section to your output titled **Estimated Marginal Means** containing a table for each main effect or interaction in your model. The table will contain a row for each cell within the effect. The values within each row provide the mean, standard error of the mean, and the boundaries for a 95% confidence interval around the mean for observations within that cell.

Graphing Interactions in an ANOVA. It is often useful to examine a plot of the means by condition when trying to interpret a significant interaction.

To get plot of means by condition from SPSS

- Perform a multifactor ANOVA as described above, but do not click the **OK** button to perform the analysis.
- Click the **Plots** button.
- Define all the plots you want to see.
 - To plot a main effect, move the factor to the **Horizontal Axis** box and click the **Add** button.
 - To plot a two-way interaction, move the first factor to the **Horizontal Axis** box, move the second factor to the **Separate Lines** box, and click the **Add** button.
 - To plot a three-way interaction, move the first factor to the **Horizontal Axis** box, move the second factor to the **Separate Lines** box, move the third factor to the **Separate Plots** box, and click the **Add** button.
- Click the **Continue** button.
- Click the **OK** button.

In addition to the standard ANOVA output, the plots you requested will appear in a section titled **Profile Plots**.

Post-hoc comparisons for when you have two or more factors. Graphing the means from a two-way or three-way between-subject ANOVA shows you the basic form of the significant interaction. However, the analyst may also wish to perform post-hoc analyses to determine which means differ from one another. If you want to compare the levels of a single factor to one another, you can follow the post-hoc procedures described in the section on one-way ANOVA. Comparing the individual cells formed by the combination of two or more factors, however, is slightly more complicated. SPSS does provide options to directly make such comparisons. Fortunately, there is a very easy method that allows one to perform post-hocs comparing all cell means to one another within a between-subjects interaction.

We will work with a specific example to illustrate how to perform this analysis in SPSS. Suppose that you wanted to compare all of the means within a 2x2x3 between-subjects factorial design. The basic idea is to create a new variable that has a different value for each cell in the above design, and then use the post-hoc procedures available in one-way ANOVA to perform your comparisons. The total number of cells in an interaction can be determined by multiplying together the number of levels in each factor composing the interaction. In our example, this would mean that our new variable would need to have $2*2*3=12$ different levels, each corresponding to a unique combination of our three IVs.

One way to create this variable would be to use the **Recode** function described above. However, there is an easier way to do this if your IVs all use numbers to code the different levels. In our example we will assume that the first factor (A) has two levels coded by the values 1 and 2, the second factor (B) has two levels again coded by the values 1 and 2, and that the third factor (C) has three levels coded by the values 1, 2, and 3. In this case, you can use the **Compute** function to calculate your new variable using the formula:

$$\text{newcode} = (\text{A} * 100) + (\text{B} * 10) + \text{C}$$

In this example, **newcode** would always be a three-digit number. The first digit would be equal to the level on variable A, the second digit would be equal to the level on variable B, while the third digit would be equal to the level on variable C. There are two benefits to using this transformation. First, it can be completed in a single step, whereas assigning the groups manually would take several separate steps. Second, you can directly see the correspondence between the levels of the original factors and the level of the composite variable by looking at the digits of the composite variable. If you actually used the values of 1 through 12 to represent the different cells in your new variable, you would likely need to reference a table to know the relationships between the values of the composite and the values of the original variables. If you ever want to create a composite of a different number of factors (besides 3 factors, like in this example), you follow the same general principle, basically multiplying each factor by decreasing powers of 10, such as the following examples.

$\text{newcode} = (A \times 10) + B$ (for a two-way interaction)

$\text{newcode} = (A \times 1000) + (B \times 100) + (C \times 10) + D$ (for a four-way interaction)

Regardless of which procedure you use to create the composite variable, you would perform the post-hoc in SPSS by taking the following steps.

- Choose **Analyze** → **General Linear Model** → **Univariate**.
- Move the DV to the **Dependent Variable** box.
- Move the composite variable to the **Fixed Factor(s)** box.
- Click the **Post-Hoc** button.
- Move the composite variable to the **Post-Hoc Tests for** box.
- Check the boxes next to the post-hoc tests you want to perform.
- Click the **Continue** button.
- Click the **OK** button.

The post-hoc analyses will be reported in the **Multiple Comparisons** and **Homogenous Subsets** sections, as described above under one-way between-subjects ANOVA.

One-way within-subjects ANOVA

A one-way within-subjects ANOVA allows you to determine if there is a relationship between a categorical IV and a continuous DV, where each subject is measured at every level of the IV. Within-subject ANOVA should be used whenever want to compare 3 or more groups where the same subjects are in all of the groups. To perform a within-subject ANOVA in SPSS you must have your data set organized so that the subject is the unit of analysis and you have different variables containing the value of the DV at each level of your within-subjects factor.

To perform a within-subject ANOVA in SPSS:

- Choose **Analyze** → **General linear model** → **Repeated measures**.
- Type the name of the factor in the **Within-Subjects Factor Name** box.
- Type the number of groups the factor represents in the **Number of Levels** box.
- Click the **Add** button.
- Click the **Define** button.
- Move the variables representing the different levels of the within-subjects factor to the **Within-Subjects Variables** box.

- Click the **OK** button.

The output of this analysis will contain the following sections.

- **Within-Subjects Factors.** Tells you what variables represent the different levels of your factor.
- **Multivariate Tests.** Contains the first test of your within-subject factor, making use of multivariate analysis. Multivariate analysis actually provides a matrix of results, which would naturally be very difficult to interpret on its own. Statisticians have therefore developed four common methods of converting the results of a multivariate test to an F-test. The most commonly used and accepted statistic is *Wilk's Lambda*. More recently statisticians have used the *Pillai-Bartlett Trace*, since research has indicated that this statistic is somewhat more robust to violations of the model assumptions than Wilk's lambda. It is therefore recommended that you base your conclusions on one of these two statistics. The decision is often moot, however, since the different statistics almost always produce similar F conversions. When you report your results in a paper, you typically state the method you used, provide the resulting F-statistic, its degrees of freedom, and its p-value.
- **Mauchly's Test of Sphericity.** The second way of testing your within-subjects factor is called *repeated measures*. This method more powerful than multivariate analysis but makes the additional assumption that the correlations between your within-subjects levels are all the same. This table provides a test of this, also called the assumption of *sphericity*. A significant test means that sphericity has been violated, indicating that you should not use the uncorrected results of a repeated-measures analysis. You can either use the multivariate results, or you can apply a correction for the violation to the repeated-measures results. If your within-subjects factor only has two levels you will always have perfect sphericity.
- **Tests of Within-Subjects Effects.** Contains the results of a repeated-measures test of your within-subjects factor. If the assumption of sphericity is satisfied, you examine the test provided in the row labeled *Sphericity Assumed*. If the sphericity assumption is violated, you should examine the tests provided in either the row titled *Greenhouse-Geisser* or *Huynh-Feldt*, which provide tests corrected for your assumption violations. If you observe a significant effect, this indicates (like in between-subjects ANOVA) that there is a difference somewhere among the means. As with between-subjects ANOVA, post-hoc comparisons are required to pinpoint which means are different from each other.
- **Tests of Within-Subjects Contrasts.** Provides the results of polynomial contrasts among your within-subject conditions. These can provide you with some information about the specific nature of the relation between your factor and the DV. However, the results will be meaningless unless your groups have some type of ordinal relation with each other.
- **Tests of Between-Subjects Effects.** This section is not typically examined when performing a one-way within-subjects ANOVA.

You can ask SPSS to provide you with the means within each level of your within-subjects factor by clicking the **Options** button in the variable selection window and moving your within-subjects variable to the **Display Means For** box. This will add a section to your output titled **Estimated Marginal Means** containing a table with a row for each level of your factor. The

values within each row provide the mean, standard error of the mean, and the boundaries for a 95% confidence interval around the mean for observations within that cell.

Multifactor within-subjects ANOVA

Just as you can use ANOVA to examine multiple between-subjects factors, so can you use it to examine multiple within-subjects factors. Multifactor ANOVA can determine the independent influence of each of your IVs on the DV (main effects) as well as the extent to which the effect of an IV on your DV depends on the level of other IVs in your model (interactions).

To perform an ANOVA with two or more within-subject factors in SPSS

- Choose **Analyze** → **General linear model** → **Repeated measures**.
- Next you define the within-subject factor(s). For each factor
 - Enter the name of the factor in the **Within-Subject Factor Name** box.
 - Enter the number of levels the factor has in the **Number of Levels** box.
 - Click the **Add** button.
- Click the **Define** button.
- The next thing you will need to do is identify which variables correspond to the particular combinations of your within-subject factors. The **Within-Subject Variables** box in the next window will contain something that looks like the following.

```

__?__ (1,1)
__?__ (1,2)
__?__ (2,1)
__?__ (2,2)
__?__ (3,1)
__?__ (3,2)

```

The numbers on the right-hand side correspond to the levels of your factors. The first number corresponds to the level of your first factor, while the second number corresponds to the level of your second factor. The order of your factors will be listed above the **Within-Subjects Variables** box. For each combination represented in this box you should select the corresponding variable in the box on the left-hand side and then press the arrow button next to the **Within-Subjects Variables** box. It doesn't matter what level of a variable you decide to associate with a number on this screen, but you must be sure that you are consistent. For example, the variable you put in the (3,1) slot should have the same level of the first factor as the variable in the (3,2) slot, and the variable you put in the (1,2) slot should have the same level of the second factor as the variable you put in the (2,2) slot.

- Click the **OK** button.

The output of this analysis will include the following sections.

- **Within-Subjects Factors**. Tells you what variables represent each combination of your factors.
- **Multivariate Tests**. Contains the first test of your main effects and interactions, making use of multivariate analysis. The test statistics provided here can be interpreted in the same way as described in the *One-way within-subjects ANOVA* section. A significant main effect indicates that at least two of the groups composing that factor have

significantly different means. A significant interaction between a set of factors indicates that the influence of any one factor involved in the interaction significantly changes under different levels of the other factors in the interaction.

- **Mauchly's Test of Sphericity.** Provides a test of the sphericity assumption for each of your within-subject terms (including both main effects and interactions).
- **Tests of Within-Subjects Effects.** Contains the repeated-measures tests of your within-subjects terms. If the assumption of sphericity is satisfied, you examine the test provided in the row labeled *Sphericity Assumed*. If the sphericity assumption is violated, you should examine the tests provided in either the row titled *Greenhouse-Geisser* or *Huynh-Feldt*, which provide tests corrected for your assumption violations.
- **Tests of Within-Subjects Contrasts.** Provides the results of polynomial contrasts among your within-subject conditions. These will have no meaning unless your levels have an ordinal relationship.
- **Tests of Between-Subjects Effects.** This section is not typically examined when performing a multifactor within-subjects ANOVA.

You can ask SPSS to provide you with the means within the levels of your main effects or your interactions by clicking the **Options** button in the variable selection window and moving the appropriate term to the **Display Means For** box. This will add a section to your output titled **Estimated Marginal Means** containing a table for each main effect or interaction in your model. The table will contain a row for each cell within the effect. The values within each row provide the mean, standard error of the mean, and the boundaries for a 95% confidence interval around the mean for observations within that cell.

Post-hoc comparisons involving within-subject factors. SPSS does not provide any options that allow you to easily compare the different levels of a within-subject factor. However, there is a relatively easy way to do this. Recall that the different levels of your within-subject factor will be stored in different variables in SPSS. If you want to see if there is a difference between two particular levels of a within-subject factor, you can create a new variable that is a difference between the two variables corresponding to the different levels you want to compare. To see if there is a significant difference, all you need to do is test whether the mean of the difference variable is significantly different from zero. This method is completely valid, and is called a *modern within-subject contrast*. You can apply a Bonferroni correction to prevent inflation of your Type I error by dividing the alpha of each contrast by the total number of post-hoc contrasts you perform from the same analysis.

Mixed ANOVA

Mixed ANOVA allow you to simultaneously examine the effect of within-subjects and between-subjects factors within the same experiment. It allows you to detect all of the following types of effects.

- Main effects of between-subjects factors
- Main effects of within-subjects factors
- Interactions involving between-subjects factors
- Interactions involving within-subjects factors
- Interactions involving both between-subjects and within-subjects factors

To perform a mixed ANOVA in SPSS

- Choose **Analyze** → **General linear model** → **Repeated measures**.
- Next you define the within-subject factor(s). For each factor
 - Enter the name of the factor in the **Within-Subject Factor Name** box.
 - Enter the number of levels the factor has in the **Number of Levels** box.
 - Click the **Add** button.
- Click the **Define** button.
- Identify which variables are associated with each combination of your within-subjects conditions as described above in the *Multifactor within-subjects ANOVA* section.
- Move any between-subjects IVs to the **Between-subjects factor(s)** box.
- Click the **OK** button.

The output from this analysis will contain the following sections.

- **Within-Subjects Factors.** Tells you what variables represent each combination of your within-subjects factors.
- **Between-Subjects Factors.** Lists how many subjects are in the combination of each of your between-subjects factors.
- **Multivariate Tests.** Contains multivariate tests of the main effects of your within-subjects factors as well as interactions that contain at least one within-subjects factor. The test statistics provided here can be interpreted in the same way as described in the *One-way within-subjects ANOVA* section. A significant main effect indicates that at least two of the groups composing that factor have significantly different means. A significant interaction between a set of factors indicates that the influence of any one factor involved in the interaction significantly changes under different levels of the other factors in the interaction.
- **Mauchly's Test of Sphericity.** Provides a test of the sphericity assumption for each of your within-subject terms (including the main effects of your within-subjects factors as well as interactions involving at least one within-subjects factor).
- **Tests of Within-Subjects Effects.** Contains repeated-measures tests of the main effects of your within-subjects factors as well as interactions that contain at least one within-subjects factor. If the assumption of sphericity is satisfied, you examine the test provided in the row labeled *Sphericity Assumed*. If the sphericity assumption is violated, you should examine the tests provided in either the row titled *Greenhouse-Geisser* or *Huynh-Feldt*, which provide tests corrected for your assumption violations.
- **Tests of Within-Subjects Contrasts.** Provides the results of polynomial contrasts among your within-subject main effects and interactions. These will have no meaning unless your levels have an ordinal relationship.
- **Tests of Between-Subjects Effects.** Contains tests of the main effects of your between-subjects factors as well as tests of any interactions that only involve between-subjects factors.

You can ask SPSS to provide you with the means within the levels of your main effects or your interactions (whether they involve within-subjects factors, between-subjects factors, or both) by clicking the **Options** button in the variable selection window and moving the appropriate term to the **Display Means For** box. This will add a section to your output titled **Estimated Marginal Means** containing a table for each main effect or interaction in your model. The table will

contain a row for each cell within the effect. The values within each row provide the mean, standard error of the mean, and the boundaries for a 95% confidence interval around the mean for observations within that cell.

MANOVA

MANOVA (multivariate analysis of variance) is a statistical procedure that allows you to determine if a set of categorical predictor variables can explain the variability in a set of continuous response variables. It is also possible to include continuous predictor variables either as covariates or as true independent variables in the design (so that you can test for the effect of interactions).

MANOVA is related to within-subject ANOVA in that both of these analyses examine multiple measurements from each case (i.e., participant) in your data set. Whether you should perform a MANOVA or a within-subject ANOVA depends on the relationship between the measurements. If the different measurements reflect observations at different levels of a theoretical factor, then you should perform a within-subject ANOVA. For example, you might look at a person's heart rate over successive days, such that the different measurements represent different levels of a "time" factor. If the measurements instead reflect different dependent variables, then you should perform a MANOVA. For example, using MANOVA you could simultaneously test whether a treatment program affects participants' responses on a depression scale, their GPA, and their performance on a reaction-time task.

The primary purpose of MANOVA is to show that an independent variable (manipulated either within- or between-subjects) has an overall effect on a collection of continuous dependent variables. If you have a large number of dependent variables, you can perform a MANOVA to see if there is any effect of your independent variables, taking into account the number of different dependent variables you are examining.

If one had multiple dependent variables, he/she could perform an ANOVA on each to examine the effect of the independent variable. However, if one were concerned that performing these multiple tests would increase the Type I error rate, a MANOVA would be a useful alternative, as it is a single test of the independent variable's influence on the collection of dependent variables. In other words, MANOVA can act as protection against an inflation of your Type I error rate from performing a large number of analyses investigating the same hypothesis. If there is a significant effect of the independent variable in the MANOVA, one could then follow up that MANOVA with univariate ANOVAs (ANOVAs with a single dependent variable). Simulations performed by Hummel and Sligo (1971, *Psychological Bulletin*) and Rencher and Scott (1990, *Communications in Statistics: Simulation and Computation*) have demonstrated that the overall experimentwide error rate when you follow up a significant multivariate test with univariate analyses is almost always below the established alpha. If the univariate tests are performed without consideration of ways to protect against alpha inflation, however, there is a significant increase in the experimentwide error rate. The most common alternative to a multivariate test would be the application of a Bonferroni correction, where the experimenter divides the alpha for each individual test by the number of tests. However, simulations demonstrate that this method leads to an overall experimentwide error rate that is substantially below the established alpha. Many researchers feel that the use of MANOVA is the best alternative since it provides good protection against alpha inflation and is more powerful than applying a Bonferroni correction. However, it must be noted that this method does not guarantee that the experimentwide error rate

will not exceed the established alpha. Most of the time the experimentwide error rate will be below alpha, but it will occasionally exceed it slightly.

Performing a MANOVA is *not* the same thing as looking for an effect on the average of your dependent variables. Therefore, it is also different from looking for a main effect of a between-subjects variable within a repeated measures analysis. One common misconception is that you cannot use MANOVA if the effect of your independent variable on the dependents varies in terms of direction, because the effects will cancel each other out. In truth, the dependent variables are never combined together in this way. MANOVA separately considers the effect of your independent variables on your dependents. It actually produces a matrix of results, which separately contains the influence of your independents on each of your dependent variables¹. The multivariate test of an independent variable does not require that it affect each dependent variable in the same way. What is important is just the extent to which your independent variables create differences in each of your dependent variables. For example, reverse coding one of the dependent variables would have absolutely no influence on a MANOVA.

Performing a MANOVA using interactive mode

You can use the interactive mode of SPSS to perform a MANOVA if all of your independent variables are manipulated between subjects.

To perform a MANOVA using the interactive mode in SPSS

- Choose **Analyze** → **General Linear Model** → **Multivariate**.
- Move the DVs you want to examine to the **Dependent Variables** box.
- Move any categorical IVs to the **Fixed Factor(s)** box.
- Move any continuous IVs to the **Covariate(s)** box.
- By default, SPSS will build a model including all interactions between the categorical independent variables, but no interactions with the continuous independent variables. To analyze a different model you must take the following steps.
 - Click the **Model** button.
 - Click the radio button next to **Custom**.
 - Add all of main effects to your model by clicking the IVs in the box labeled **Factors and Covariates**, setting the pull-down menu to **Main effects**, and clicking the arrow button.
 - Add each of the interaction terms to your model. You can do this one at a time by selecting the variables included in the interaction in the box labeled **Factors and Covariates**, setting the pull-down menu to **Interaction**, and then clicking the arrow button. You can also use the setting on the pull-down menu to tell SPSS to add all possible 2-way, 3-way, 4-way, or 5-way interactions that can be made between the selected variables to your model. You should be sure to center any continuous variables that you want to interact with other variables.
 - Click the **Continue** button.

¹ Actually, what MANOVA does is determine the effect of your independent variables on the principle components that can be calculated from your dependent variables. However, thinking of it the way described in the text is a little simpler and is basically accurate, since the components represent the dimensions of variability found in your dependent variables.

- You can ask SPSS to perform post-hoc contrasts by clicking the **Contrasts** button. SPSS reports the contrasts separately for each dependent variable but will *not* produce a multivariate contrast testing for a difference between two groups across all of the dependent variables. You can also ask SPSS to create new variables to hold the predicted values, residuals, or diagnostics from your model by clicking the **Save** button. If you ask SPSS to save one of these values, you will get a number of new variables in your data set equal to the total number of dependent variables in your model. These values will all be based on the univariate ANOVAs predicting the individual dependent variables from your independent variables.
- Click the **Ok** button when you are ready for SPSS to perform the analysis.

Performing a MANOVA using syntax

If you want to perform a MANOVA including one or more within-subjects factors (so that you measure each of your DVs under each combination of your within-subjects factors), you will need to conduct your analysis using SPSS syntax. You will need to have a variable in your data set for the measurement of each dependent variable at each combination of the within-subjects factors in your study. For example, if you had a within-subjects factor with two levels and three dependent variables, you should have six different response variables in your data set.

Below is an example of the code that would be used to analyze a design with one within-subject factor with two levels, one within-subject factor with three levels, and one between-subjects factor with two levels, and four dependent variables. In this case we need a total of 4 (DVs) X 2 (levels of **wthn1**) X 3 (levels of **wthn2**) = 24 different variables. We will assume that the variable names start with **var1**, **var2**, or **var3**, depending on which dependent measure it reflects. The variable names end with **c11**, **c12**, **c13**, **c21**, **c22**, or **c23**, indicating the condition in which the measurement was taken. This way, you can tell exactly what each variable measures just by looking at its name. We also have a variable called **betwn** coding the between-subjects condition for each case.

GLM

```
var1c11 var1c12 var1c13 var1c21 var1c22 var1c23 var2c11 var2c12
var2c13 var2c21 var2c22 var2c23 var3c11 var3c12 var3c13 var3c21
var3c22 var3c23 var4c11 var4c12 var4c13 var4c21 var4c22 var4c23
BY betwn
  /WSFACTOR = wthn1 2 wthn2 3
  /METHOD = SSTYPE(3)
  /CRITERIA = ALPHA(.05)
  /WSDESIGN = wthn1 wthn2 wthn1*wthn2
  /DESIGN = betwn .
```

As you can see, MANOVA is performed using the GLM statement. The first thing you do is list out all of the variables that contain the measurements on your dependent variables at each combination of your within-subject factors. You must list out the variables in a very particular way, so that SPSS knows what variables correspond to the same dependent variables, and which correspond to the same levels of your within-subject factors. As you can see in the example, the variables must first be grouped by the dependent variable. Within each dependent variable, you then group them by the levels of the first within-subject factor. Finally, within each level of the first within-subject factor, you organize the variables by the levels of the second within-subject

factor. In this way, the dependent variable changes the slowest, the first within-subject factor changes the second slowest, and the second within-subject factor changes the fastest. This list is then followed by the word **BY**, which in turn is followed by a list between-subjects factors.

After the variable list, you issue the subcommand **/WSFACTOR** to identify the names and number of levels of your within-subject factors. **/METHOD** establishes which sums of squares you want to base your tests on, and **/CRITERIA** sets the experimentwide alpha level. Following this you issue the **/WSDESIGN** subcommand where you list what main effects and interactions between your within-subject factors you want included in your model. The **/DESIGN** subcommand does the same thing, except for between-subjects factors. Your model will also include all the interactions between the terms you define in the **/WSDESIGN** and **/DESIGN** subcommands. In this example the model would also include the two-way interactions **betwn*wthn1**, **betwn*wthn2**, as well as the three-way interaction **betwn*wthn1*wthn2**.

Interpreting MANOVA output

Performing a MANOVA in SPSS will produce the following sections of output.

- **Within-subject Factors.** Describes how SPSS assigns the different variables to the different measures and within-subject conditions. You should always make sure that SPSS is interpreting your variables in the correct way if you have within-subject factors. It is very easy to make an error when listing out the variables in the first part of the GLM command, which would completely throw off your analyses.
- **Between-Subjects Factors.** Reports the between-subjects conditions in your design and the number of subjects in each of those cells.
- **Multivariate Tests.** Provides the results of the multivariate tests of each effect in your model. As mentioned above, MANOVA actually produces a matrix of results. However, these results are very difficult to interpret on their own, and so they are typically converted to an F statistic to make the determination of the p-value easier. There are four common methods of converting the results of a MANOVA to an F. SPSS reports all four values as well as the corresponding F statistics and degrees of freedom. The most commonly used and accepted statistic is Wilk's Lambda. More recently statisticians have used the Pillai-Bartlett trace, since research has indicated that this statistic is somewhat more robust to violations of the model assumptions than Wilk's lambda. It is therefore recommended that you base your conclusions on one of these two statistics. The decision is often moot, however, since the different statistics almost always produce similar F conversions. When you report your results in a paper, you typically state the method you used and provide the resulting F-statistic with its degrees of freedom and p-value.
- **Mauchly's Test of Sphericity.** Tells you whether your various dependent variables meet the assumption for a repeated measures ANOVA. This table only appears if you have a within-subject factor in the design. You will see a sphericity test for each dependent variable in the design. If a sphericity test is significant, it means that this assumption is violated, so you should NOT interpret the results of a repeated-measures analysis on that dependent variable without correction. This has nothing to do with the overall MANOVA itself, but just the follow-up repeated measures analyses on the individual DVs.
- **Multivariate.** Reports the ability of your within-subject effects to account for variability in the average of your dependent variables in a multivariate analysis. This test will not be

meaningful unless your measures are all on the same scale of measurement and coded in the same direction.

- **Univariate Tests.** Reports the ability of your within-subject effects to account for variability in each of your dependent variables individually in a repeated measures analysis. For more information on interpreting this table, refer to the *Multifactor within-subjects ANOVA* section of these notes.
- **Tests of Within-Subjects Contrasts.** Provides the results of polynomial contrasts among your within-subject main effects and interactions. These will have no meaning unless your levels have an ordinal relationship.
- **Tests of Between-Subjects Effects.** which reports the ability of each of your between-subjects effects to account for variability in each of your dependent measures individually. For more information on this table, refer to the *Multifactor between-subjects ANOVA* section of these notes.

CORRELATION

Pearson correlation

A Pearson correlation measures the strength of the linear relationship between two continuous variables. A linear relationship is one that can be captured by drawing a straight line on a scatterplot between the two variables of interest. The value of the correlation provides information both about the nature and the strength of the relationship.

- Correlations range between -1.0 and 1.0.
- The sign of the correlation describes the direction of the relationship. A positive sign indicates that as one variable gets larger the other also tends to get larger, while a negative sign indicates that as one variable gets larger the other tends to get smaller.
- The magnitude of the correlation describes the strength of the relationship. The further that a correlation is from zero, the stronger the relationship is between the two variables. A zero correlation would indicate that the two variables aren't related to each other at all.

Correlations only measure the strength of the *linear* relationship between the two variables. Sometimes you have a relationship that would be better measured by a curve of some sort rather than a straight line. In this case the correlation coefficient would not provide a very accurate measure of the strength of the relationship. If a line accurately describes the relationship between your two variables, your ability to predict the value of one variable from the value of the other is directly related to the correlation between them. When the points in your scatterplot are all clustered closely about a line your correlation will be large and the accuracy of the predictions will be high. If the points tend to be widely spread your correlation will be small and the accuracy of your predictions will be low.

The Pearson correlation assumes that both of your variables have normal distributions. If this is not the case then you might consider performing a Spearman rank-order correlation instead (described below).

To perform a Pearson correlation in SPSS

- Choose **Analyze** → **Correlate** → **Bivariate**.
- Move the variables you want to correlate to the **Variables** box.
- Click the **OK** button.

The output of this analysis will contain the following section.

- **Correlations.** This section contains the correlation matrix of the variables you selected. A variable always has a perfect correlation with itself, so the diagonals of this matrix will always have values of 1. The other cells in the table provide you with the correlation between the variable listed at the top of the column and the variable listed to the left of the row. Below this is a p-value testing whether the correlation differs significantly from zero. Finally, the bottom value in each box is the sample size used to compute the correlation.

Point-biserial correlation

The point-biserial correlation captures the relationship between a dichotomous (two-value) variable and a continuous variable. If the analyst codes the dichotomous variable with values of

0 and 1, and then computes a standard Pearson correlation using this variable, it is mathematically equivalent to the point-biserial correlation. The interpretation of this variable is similar to the interpretation of the Pearson correlation. A positive correlation indicates that group associated with the value of 1 has larger values than the group associated with the value of 0. A negative correlation indicates that group associated with the value of 1 has smaller values than the group associated with the value of 0. A value near zero indicates no relationship between the two variables.

To perform a point-biserial correlation in SPSS

- Make sure your categories are indicated by values of 0 and 1.
- Obtain the Pearson correlation between the categorical variable and the continuous variable, as discussed above.

The result of this analysis will include the same sections as discussed in the *Pearson correlation* section.

Spearman rank correlation

The Spearman rank correlation is a nonparametric equivalent to the Pearson correlation. The Pearson correlation assumes that both of your variables have normal distributions. If this assumption is violated for either of your variables then you may choose to perform a Spearman rank correlation instead. However, the Spearman rank correlation is a less powerful measure of association, so people will commonly choose to use the standard Pearson correlation even when the variables you want to consider are moderately nonnormal. The Spearman Rank correlation is typically preferred over Kenda's tau, another nonparametric correlation measure, because its scaling is more consistent with the standard Pearson correlation.

To perform a Spearman rank correlation in SPSS

- Choose **Analyze** → **Correlate** → **Bivariate**.
- Move the variables you want to correlate to the **Variables** box.
- Check the box next to **Spearman**.
- Click the **OK** button.

The output of this analysis will contain the following section.

- **Correlations.** This section contains the correlation matrix of the variables you selected. The Spearman rank correlations can be interpreted in exactly the same way as you interpret a standard Pearson correlation. Below each correlation SPSS provides a p-value testing whether the correlation is significantly different from zero, and the sample size used to compute the correlation.

REGRESSION

Regression is a statistical tool that allows you to predict the value of one continuous variable from one or more other variables. When you perform a regression analysis, you create a regression equation that predicts the values of your DV using the values of your IVs. Each IV is associated with specific coefficients in the equation that summarizes the relationship between that IV and the DV. Once we estimate a set of coefficients in a regression equation, we can use hypothesis tests and confidence intervals to make inferences about the corresponding parameters in the population. You can also use the regression equation to predict the value of the DV given a specified set of values for your IVs.

Simple Linear Regression

Simple linear regression is used to predict the value of a single continuous DV (which we will call Y) from a single continuous IV (which we will call X). Regression assumes that the relationship between IV and the DV can be represented by the equation

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where Y_i is the value of the DV for case i , X_i is the value of the IV for case i , β_0 and β_1 are constants, and ε_i is the error in prediction for case i . When you perform a regression, what you are basically doing is determining estimates of β_0 and β_1 that let you best predict values of Y from values of X. You may remember from geometry that the above equation is equivalent to a straight line. This is no accident, since the purpose of simple linear regression is to define the line that represents the relationship between our two variables. β_0 is the intercept of the line, indicating the expected value of Y when $X = 0$. β_1 is the slope of the line, indicating how much we expect Y will change when we increase X by a single unit.

The regression equation above is written in terms of population parameters. That indicates that our goal is to determine the relationship between the two variables in the population as a whole. We typically do this by taking a sample and then performing calculations to obtain the estimated regression equation

$$Y_i = b_0 + b_1 X_i$$

Once you estimate the values of b_0 and b_1 , you can substitute in those values and use the regression equation to predict the expected values of the DV for specific values of the IV. Predicting the values of Y from the values of X is referred to as regressing Y on X. When analyzing data from a study you will typically want to regress the values of the DV on the values of the IV. This makes sense since you want to use the IV to explain variability in the DV. We typically calculate b_0 and b_1 using *least squares estimation*. This chooses estimates that minimize the sum of squared errors between the values of the estimated regression line and the actual observed values.

In addition to using the estimated regression equation for prediction, you can also perform hypothesis tests regarding the individual regression parameters. The slope of the regression equation (β_1) represents the change in Y with a one-unit change in X. If X predicts Y, then as X

increases, Y should change in some systematic way. You can therefore test for a linear relationship between X and Y by determining whether the slope parameter is significantly different from zero.

When using performing linear regression, we typically make the following assumptions about the error terms ε_i .

1. The errors have a normal distribution.
2. The same amount of error in the model is found at each level of X.
3. The errors in the model are all independent.

To perform a simple linear regression in SPSS

- Choose **Analyze** → **Regression** → **Linear**.
- Move the DV to the **Dependent** box.
- Move the IV to the **Independent(s)** box.
- Click the **Continue** button.
- Click the **OK** button.

The output from this analysis will contain the following sections.

- **Variables Entered/Removed**. This section is only used in model building and contains no useful information in simple linear regression.
- **Model Summary**. The value listed below **R** is the correlation between your variables. The value listed below **R Square** is the proportion of variance in your DV that can be accounted for by your IV. The value in the **Adjusted R Square** column is a measure of model fit, adjusting for the number of IVs in the model. The value listed below **Std. Error of the Estimate** is the standard deviation of the residuals.
- **ANOVA**. Here you will see an ANOVA table, which provides an F test of the relationship between your IV and your DV. If the F test is significant, it indicates that there is a relationship.
- **Coefficients**. This section contains a table where each row corresponds to a single coefficient in your model. The row labeled **Constant** refers to the intercept, while the row containing the name of your IV refers to the slope. Inside the table, the column labeled **B** contains the estimates of the parameters and the column labeled **Std. Error** contains the standard error of those parameters. The column labeled **Beta** contains the standardized regression coefficient, which is the parameter estimate that you would get if you standardized both the IV and the DV by subtracting off their mean and dividing by their standard deviations. Standardized regression coefficients are sometimes used in multiple regression (discussed below) to compare the relative importance of different IVs when predicting the DV. In simple linear regression, the standardized regression coefficient will always be equal to the correlation between the IV and the DV. The column labeled **t** contains the value of the t-statistic testing whether the value of each parameter is equal to zero. The p-value of this test is found in the column labeled **Sig**. If the value for the IV is significant, then there is a relationship between the IV and the DV. Note that the square of the t statistic is equal to the F statistic in the ANOVA table and that the p-values of the two tests are equal. This is because both of these are testing whether there is a significant linear relationship between your variables.

Multiple Regression

Sometimes you may want to explain variability in a continuous DV using several different continuous IVs. Multiple regression allows us to build an equation predicting the value of the DV from the values of two or more IVs. The parameters of this equation can be used to relate the variability in our DV to the variability in specific IVs. Sometimes people use the term multivariate regression to refer to multiple regression, but most statisticians do not use “multiple” and “multivariate” as synonyms. Instead, they use the term “multiple” to describe analyses that examine the effect of two or more IVs on a single DV, while they reserve the term “multivariate” to describe analyses that examine the effect of any number of IVs on two or more DVs.

The general form of the multiple regression model is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \varepsilon_i.$$

The elements in this equation are the same as those found in simple linear regression, except that we now have k different parameters which are multiplied by the values of the k IVs to get our predicted value. We can again use least squares estimation to determine the estimates of these parameters that best our observed data. Once we obtain these estimates we can either use our equation for prediction, or we can test whether our parameters are significantly different from zero to determine whether each of our IVs makes a significant contribution to our model.

Care must be taken when making inferences based on the coefficients obtained in multiple regression. The way that you interpret a multiple regression coefficient is somewhat different from the way that you interpret coefficients obtained using simple linear regression. Specifically, the value of a multiple regression coefficient represents the ability of part of the corresponding IV that is unrelated to the other IVs to predict the part of the DV that is unrelated to the other IVs. It therefore represents the unique ability of the IV to account for variability in the DV. One implication of the way coefficients are determined is that your parameter estimates become very difficult to interpret if there are large correlations among your IVs. The effect of these relationships on multiple regression coefficients is called *multicollinearity*. This changes the values of your coefficients and greatly increases their variance. It can cause you to find that none of your coefficients are significantly different from zero, even when the overall model does a good job predicting the value of the DV.

One implication of the way coefficients are determined is that your parameter estimates become very difficult to interpret if there are large correlations among your IVs. The typical effect of multicollinearity is to reduce the size of your parameter estimates. Since the value of the coefficient is based on the unique ability for an IV to account for variability in a DV, if there is a portion of variability that is accounted for by multiple IVs, all of their coefficients will be reduced. Under certain circumstances multicollinearity can also create a suppression effect. If you have one IV that has a high correlation with another IV but a low correlation with the DV, you can find that the multiple regression coefficient for the second IV from a model including both variables can be larger (or even opposite in direction!) compared to the coefficient from a model that doesn't include the first IV. This happens when the part of the second IV that is independent of the first IV has a different relationship with the DV than does the part that is

related to the first IV. It is called a suppression effect because the relationship that appears in multiple regression is suppressed when you just look at the variable by itself.

To perform a multiple regression in SPSS

- Choose **Analyze** → **Regression** → **Linear**.
- Move the DV to the **Dependent** box.
- Move all of the IVs to the **Independent(s)** box.
- Click the **Continue** button.
- Click the **OK** button.

The SPSS output from a multiple regression analysis contains the following sections.

- **Variables Entered/Removed**. This section is only used in model building and contains no useful information in standard multiple regression.
- **Model Summary**. The value listed below **R** is the multiple correlation between your IVs and your DV. The value listed below **R square** is the proportion of variance in your DV that can be accounted for by your IV. The value in the **Adjusted R Square** column is a measure of model fit, adjusting for the number of IVs in the model. The value listed below **Std. Error of the Estimate** is the standard deviation of the residuals.
- **ANOVA**. This section provides an F test for your statistical model. If this F is significant, it indicates that the model as a whole (that is, all IVs combined) predicts significantly more variability in the DV compared to a null model that only has an intercept parameter. Notice that this test is affected by the number of IVs in the model being tested.
- **Coefficients**. This section contains a table where each row corresponds to a single coefficient in your model. The row labeled **Constant** refers to the intercept, while the coefficients for each of your IVs appear in the row beginning with the name of the IV. Inside the table, the column labeled **B** contains the estimates of the parameters and the column labeled **Std. Error** contains the standard error of those estimates. The column labeled **Beta** contains the standardized regression coefficient. The column labeled **t** contains the value of the t-statistic testing whether the value of each parameter is equal to zero. The p-value of this test is found in the column labeled **Sig**. A significant t-test indicates that the IV is able to account for a significant amount of variability in the DV, independent of the other IVs in your regression model.

Multiple regression with interactions

In addition to determining the independent effect of each IV on the DV, multiple regression can also be used to detect *interactions* between your IVs. An interaction measures the extent to which the relationship between an IV and a DV depends on the level of other IVs in the model. For example, if you have an interaction between two IVs (called a two-way interaction) then you expect that the relationship between the first IV and the DV will be different across different levels of the second IV. Interactions are symmetric, so if you have an interaction such that the effect of IV1 on the DV depends on the level of IV2, then it is also true that the effect of IV2 on the DV depends on the level of IV1. It therefore does not matter whether you say that you have an interaction between IV1 and IV2 or an interaction between IV2 and IV1. You can also have interactions between more than two IVs. For example, you can have a three-way interaction between IV1, IV2, and IV3. This would mean that the two-way interaction between IV1 and IV2 depends on the level of IV3. Just like two-way interactions, three-way interactions are also

independent of the order of the variables. So the above three-way interaction would also mean that the two-way interaction between IV1 and IV3 is dependent on the level of IV2, and that the two-way interaction between IV2 and IV3 depends on the level of IV1.

It is possible to have both main effects and interactions at the same time. For example, you can have a general trend that the value of the DV increases when the value of a particular IV increases along with an interaction such that the relationship is stronger when the value of a second IV is high than when the value of that second IV is low. You can also have lower order interactions in the presence of a higher order interaction. Again, the lower-order interaction would represent a general trend that is modified by the higher-order interaction.

You can use linear regression to determine if there is an interaction between a pair of IVs by adding an interaction term to your statistical model. To detect the interaction effect of two IVs (X_1 and X_2) on a DV (Y) you would use linear regression to estimate the equation

$$Y_i = b_0 + b_1X_{i1} + b_2X_{i2} + b_3X_{i1}X_{i2}.$$

You construct the variable for the interaction term $X_{i1}X_{i2}$ by literally multiplying the value of X_1 by the value of X_2 for each case in your data set. If the test of b_3 is significant, then the two predictors have an interactive effect on the outcome variable.

In addition to the interaction term itself, your model must contain all of the main effects of the variables involved in the interaction as well as all of the lower-order interaction terms that can be created using those main effects. For example, if you want to test for a three-way interaction you must include the three main effects as well as all of the possible two-way interactions that can be made from those three variables. If you do not include the lower-order terms then the test on the highest order interaction will produce incorrect results.

It is important to *center* the variables that are involved in an interaction before including them in your model. That is, for each independent variable, the analyst should subtract the mean of the independent variable from each participant's score on that variable. The interaction term should then be constructed from the centered variables by multiplying them together. The model itself should then be tested using the centered main effects and the constructed interaction term. Centering your independent variables will not change their relationship to the dependent variable, but it will reduce the collinearity between the main effects and the interaction term. If the variables are not centered then none of the coefficients on terms involving IVs involved in the interaction will be interpretable except for the highest-order interaction. When the variables are centered, however, then the coefficients on the IVs can be interpreted as representing the main effect of the IV on the DV, averaging over the other variables in the interaction. The coefficients on lower-order interaction terms can similarly be interpreted as the testing the average strength of that lower-order interaction, averaging over the variables that are excluded from the lower-order interaction but included in the highest-order interaction term. Centering has the added benefit of reducing the collinearity between the main effect and interaction terms.

You can perform a multiple regression including interaction terms in SPSS just like you would a standard multiple regression if you create your interaction terms ahead of time. However,

creating these variables can be tedious when analyzing models that contain a large number of interaction terms. Luckily, if you choose to analyze your data using the **General Linear Model** procedure, SPSS will create these interaction terms for you (although you still need to center all of your original IVs beforehand). To analyze a regression model this way in SPSS

- Center the IVs involved in the interaction.
- Choose **Analyze → General Linear Model → Univariate**.
- Move your DV to the box labeled **Dependent Variable**.
- Move all of the main effect terms for your IVs to the box labeled **Covariate(s)**.
- Click the **Options** button.
- Check the box next to **Parameter estimates**. By default this procedure will only provide you with tests of your IVs and not the actual parameter estimates.
- Click the **Continue** button.
- By default SPSS will not include interactions between continuous variables in its statistical models. However, if you build a custom model you can include whatever terms you like. You should therefore next build a model that includes all of the main effects of your IVs as well as any desired interactions. To do this
 - Click the **Model** button.
 - Click the radio button next to **Custom**.
 - Select all of your IVs, set the drop-down menu to **Main effects**, and click the arrow button.
 - For each interaction term, select the variables involved in the interaction, set the drop-down menu to **Interaction**, and click the arrow button.
 - If you want all of the possible two-way interactions between a collection of IVs you can just select the IVs, set the drop-down menu to **All 2-way**, and click the arrow button. This procedure can also be used to get all possible three-way, four-way, or five-way interactions between a collection of IVs by setting the drop-down menu to the appropriate interaction type.
- Click the **Continue** button.
- Click the **OK** button.

The output from this analysis will contain the same sections found in standard multiple regression. When referring to an interaction, SPSS will display the names of the variables involved in the interaction separated by asterisks (*). So the interaction between the variables RACE and GENDER would be displayed as RACE * GENDER.

So what does it mean if you obtain a significant interaction in regression? Remember that in simple linear regression, the slope coefficient (b_1) indicates the expected change in Y with a one-unit change in X. In multiple regression, the slope coefficient for X_1 indicates the expected change in Y with a one-unit change in X_1 , holding all other X values constant. Importantly, this change in Y with a one-unit change in X_1 is the same no matter what value the other X variables in the model take on. However, if there is a significant interaction, the interpretation of coefficients is slightly different. In this case, the slope coefficient for X_1 depends on the level of the other predictor variables in the model.

Polynomial regression

Polynomial regression models are used when the true relationship between a continuous predictor variable and a continuous dependent variable is a polynomial function, or when the curvilinear relationship is complex or unknown but can be approximated by a polynomial function.

A polynomial regression model with one predictor variable is expressed in the following way:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_{11} X_i^2 + \varepsilon_i$$

The predictor variable (X) should be centered (discussed in the section **Multiple regression with interactions**), or else the X and X² terms will be highly correlated and lead to severe multicollinearity. Additionally, you lose the ability to interpret the lower-order coefficients in a straightforward manner.

In the above model, the coefficient β_1 is typically called the “linear effect” coefficient and β_{11} is called the “quadratic effect” coefficient. If the estimate of the coefficient β_{11} is significantly different from zero then you have a significant quadratic effect in your data. If the highest-order term in a polynomial model is not significant, conventionally statisticians will remove that term from the model and rerun the regression.

The best way to choose the highest order polynomial is through a historical or theoretical analysis. There are certain types of relationships that are well known to be fitted by quadratic or cubic models. You might also determine that a specific type of relationship should exist because of the mechanisms responsible for the relationship between the IV and the DV. If you are building your model in an exploratory fashion, however, you can estimate how high of an order function you should use by the shape of the relationship between the DV and that IV. If your data appears to reverse p times (has p curves in the graph), you should use a function whose highest order parameter is raised to the power of p + 1. In multiple regression you can see whether you should add an additional term for an IV by examining a graph of the residuals against the IV. Again, if the relationship between the residuals and the IV appears to reverse p times, you should add terms whose highest order parameter is raised to the power of p + 1.

It is quite permissible to have more than one predictor variable represented in quadratic form in the same model. For instance:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_{11} X_{i1}^2 + \beta_{22} X_{i2}^2 + \varepsilon_i$$

is a model with two predictor variables, both with quadratic terms.

To perform a polynomial regression in SPSS

- Determine the highest order term that you will use for each IV.
- Center any IVs for which you will examine higher-order terms.
- For each IV, create new variables that are equal to your IV raised to the powers of 2 through the power of your highest order term. Be sure to use the centered version of your IV.

- Conduct a standard multiple regression including all of the terms for each IV.

Simultaneously testing categorical and continuous IVs

Both ANOVA and regression are actually based on the same set of statistical ideas, the **general linear model**. SPSS implements these functions in different menu selections, but the basic way that the independent variables are tested is fundamentally the same. It is therefore perfectly reasonable to combine both continuous and categorical predictor variables in the same model, even though people are usually taught to think of ANOVA and regression as separate types of analyses.

To perform an analysis in SPSS using the General Linear Model

- Choose **Analyze** → **General Linear Model** → **Univariate**.
- Move your DV to the box labeled **Dependent Variable**.
- Move any categorical IVs to the box labeled **Fixed Factor(s)**.
- Move any continuous IVs to the box labeled **Covariate(s)**.
- By default SPSS will include all possible interactions between your categorical IVs, but will only include the main effects of your continuous IVs. If this is not the model you want then you will need to define it by hand by taking the following steps.
 - Click the **Model** button.
 - Click the radio button next to **Custom**.
 - Add all of your main effects to the model by clicking all of the IVs in the box labeled **Factors and covariates**, setting the pull-down menu to **Main effects**, and clicking the arrow button.
 - Add each of the interaction terms to your model. You can do this one at a time by selecting the variables included in the interaction in the box labeled **Factors and covariates**, setting the pull-down menu to **Interaction**, and clicking the arrow button for each of your interactions.
 - You can also use the setting on the pull-down menu to tell SPSS to add all possible 2-way, 3-way, 4-way, or 5-way interactions that can be made between the selected variables to your model.
 - Click the **Continue** button.
- Click the **OK** button.

The SPSS output from running an analysis using the General Linear Model contains the following sections.

- **Between-Subjects Factors**. This table just lists out the different levels of any categorical variables included in your model.
- **Tests of Between-Subjects Effects**. This table provides an F test of each main effect or interaction that you included in your model. It indicates whether or not the effect can independently account for a significant amount of variability in your DV. This provides the same results as testing the change in model R^2 that you get from the test of the set of terms representing the effect.

Post-hoc comparisons in mixed models. You can ask SPSS to provide post-hoc contrasts comparing the different levels within any of your categorical predictor variables by clicking the *Contrasts* button in the variable selection window. If you want to compare the means of cells

resulting from combinations of your categorical predictors, you will need to recode them all into a single variable as described in the section **Post-hoc comparisons for when you have two or more factors**.

The easiest way to examine the main effect of a continuous independent variable is to graph its relationship to the dependent variable using simple linear regression.. You can obtain this using the following procedure:

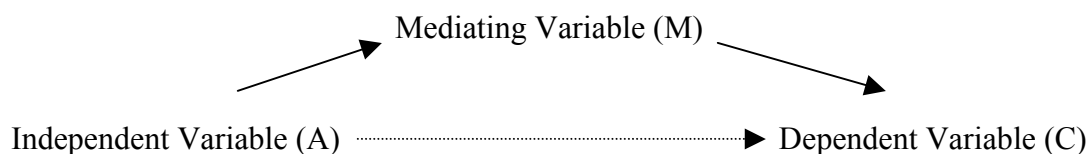
- Choose **Analyze → Regression → Curve Estimation**.
- Move your dependent variable into the **Dependent(s)** box
- Move your independent variable into the **Independent** box
- Make sure that **Plot Models** is checked
- Under the heading **Models**, make sure that only **Linear** is checked

This will produce a graph of your data along with the least-squares regression line. If you want to look at the interaction between a categorical and a continuous independent variable, you can use the **Select Cases** function (described above) to limit this graph to cases that have a particular value on the categorical variable. Using this method several times, you can obtain graphs of the relationship between the continuous variable and the dependent variable separately for each level of the categorical independent variable.

Another option you might consider would be to recode the continuous variables as categorical, separating them into groups based on their value on the continuous variables. You can then run a standard ANOVA and compare the means of the dependent variable for those high or low on the continuous variable. Even if you decide to do this, you should still base all of your conclusions on the analysis that actually treated the variable as continuous. Numerous simulations have shown that there is greater power and less error in analysis that treat truly continuous variables as continuous compared to those that analyze them in a categorical fashion.

MEDIATION

When researchers find a relationship between an independent variable (A) and a dependent variable (C), they may seek to uncover variables that mediate this relationship. That is, they may believe that the effect of variable A on variable C exists because variable A leads to a change in a mediating variable (M), which in turn effects the dependent variable (C). When a variable fully mediates a relationship, the effect of variable A on variable C disappears when controlling for the mediating variable. A variable partially mediates a relationship when the effect of variable A on variable C is significantly reduced when controlling for the mediator. A common way of expressing these patterns is the following:



You need to conduct three different regression analyses to determine if you have a mediated relationship using the traditional method

Regression 1. Predict the dependent variable (C) from the independent variable (A). The effect of the independent variable in this model must be significant. If there is no direct effect of A on C, then there is no relationship to mediate.

Regression 2. Predict the mediating variable (M) from the independent variable (A). The effect of the independent variable in this model must be significant. If the independent variable does not reliably affect the mediator, the mediator cannot be responsible for the relationship observed between A and C.

Regression 3. Simultaneously predict the value of the dependent variable (C) from both the independent variable (A) and the mediating variable (M) using multiple regression. The effect of the independent variable should be non significant (or at least significantly reduced, compared to Regression 1), whereas the effect of the mediating variable must be significant. The reduction in the relationship between A and C indicates that the mediator is accounting for a significant portion of this relationship. However, if the relationship between M and C is not significant, then you cannot clearly determine whether M mediates the relationship between A and C, or if A mediates the relationship between M and C.

One can directly test for a reduction in the effect of $A \rightarrow C$ when controlling for the mediator by performing a **Sobel Test**. This involves testing the significance of the path between A and C through M in Regression 3. While you cannot do a Sobel Test in SPSS, the website <http://www.unc.edu/~preacher/sobel/sobel.htm> will perform this for you online. If you wish to show mediation in a journal article, you will almost always be required to show the results of the Sobel Test.

CHI-SQUARE TEST OF INDEPENDENCE

A chi-square is a nonparametric test used to determine if there is a relationship between two categorical variables. Let's take a simple example. Suppose a researcher brought male and female participants into the lab and asked them which color they prefer—blue or green. The researcher believes that color preference may be related to gender. Notice that both gender (male, female) and color preference (blue, green) are categorical variables. If there is a relationship between gender and color preference, we would expect that the proportion of men who prefer blue would be different than the proportion of women who prefer blue. In general, you have a relationship between two categorical variables when the distribution of people across the categories of the first variable changes across the different categories of the second variable.

To determine if a relationship exists between gender and color preference, the chi-square test computes the distributions across the combination of your two factors that you would expect if there were no relationship between them. It then compares this to the actual distribution found in your data. In the example above, we have a 2 (gender: male, female) X 2 (color preference: green, blue) design. For each cell in the combination of the two factors, we would compute "observed" and "expected" counts. The observed counts are simply the actual number of observations found in each of the cells. The expected proportion in each cell can be determined by multiplying the marginal proportions found in a table. For example, let us say that 52% of all the participants preferred blue and 48% preferred green, whereas 40% of all of the participants were men and 60% were women. The expected proportions are presented in the table below.

Expected proportion table

	Males	Females	Marginal proportion
Blue	20.8%	31.2%	52%
Green	19.2%	28.8%	48%
Marginal proportion	40%	60%	

As you can see, you get the expected proportion for a particular cell by multiplying the two marginal proportions together. You would then determine the expected count for each cell by multiplying the expected proportion by the total number of participants in your study. The chi-square statistic is a function of the difference between the expected and observed counts across all your cells. Luckily you do not actually need to calculate any of this by hand, since SPSS will compute the expected counts for each cell and perform the chi-square test.

To perform a chi-square test of independence in SPSS

- Choose **Analyze** → **Descriptive Statistics** → **Crosstabs**.
- Put one of the variables in the **Row(s)** box
- Put the other variable in the **Column(s)** box
- Click the **Statistics** button.
- Check the box next to **Chi-square**.
- Click the **Continue** button.

- Click the **OK** button.

The output of this analysis will contain the following sections.

- **Case Processing Summary.** Provides information about missing values in your two variables.
- **Crosstabulation.** Provides you with the observed counts within each combination of your two variables.
- **Chi-Square Tests.** The first row of this table will give you the chi-square value, its degrees of freedom and the p-value associated with the test. Note that the p-values produced by a chi-square test are inappropriate if the expected count is less than 5 in 20% of the cells or more. If you are in this situation, you should either redefine your coding scheme (combining the categories with low cell counts with other categories) or exclude categories with low cell counts from your analysis.

LOGISTIC REGRESSION

The chi-square test allows us to determine if a pair of categorical variables are related. But what if you want to test a model using two or more independent variables? Most of the inferential procedures we have discussed so far require that the dependent variable be a continuous variable. The most common inferential statistics such as t-tests, regression, and ANOVA, require that the residuals have a normal distribution, and that the variance is equal across conditions. Both of these assumptions are likely to be seriously violated if the dependent variable is categorical. The answer is to use logistic regression, which does not make these assumptions and so can be used to determine the ability of a set of continuous or categorical independent variables to predict the value of a categorical dependent variable. However, standard logistic regression assumes that all of your observations are independent, so it cannot be directly used to test within-subject factors.

Logistic regression generates equations that tell you exactly how changes in your independent variables affect the probability that the observation is in a level of your dependent variable. These equations are based on predicting the *odds* that a particular observation is in one of two groups. Let us say that you have two groups: a reference group and a comparison group. The odds that an observation is in the reference group is equal to the probability that the observation is in the reference group divided by the probability that it is in the comparison group. So, if there is a 75% chance that the observation is in the reference group, the odds of it being in the reference group would be $.75/.25 = 3$. We therefore talk about odds in the same way that people do when betting at a racetrack.

In logistic regression, we build an equation that predicts the logarithm of the odds from the values of the independent variables (which is why it's called *log-istic* regression). For each independent variable in our model, we want to calculate a coefficient B that tells us what the change in the log odds would be if we would increase the value of the variable by 1. These coefficients therefore parallel those found in a standard regression model. However, they are somewhat difficult to interpret because they relate the independent variables to the log odds. To make interpretation easier, people often transform the coefficients into *odds ratios* by raising the mathematical constant e to the power of the coefficient (e^B). The odds ratio directly tells you how the odds increase when you change the value of the independent variable. Specifically, the odds of being in the reference group are multiplied by the odds ratio when the independent variable increases by 1.

One obvious limitation of this procedure is that we can only compare two groups at a time. If we want to examine a dependent variable with three or more levels, we must actually create several different logistic regression equations. If your dependent variable has k levels, you will need a total of $k-1$ logistic regression equations. What people typically do is designate a specific level of your dependent variable as the reference group, and then generate a set of equations that each compares one other level of the dependent variable to that group. You must then examine the behavior of your independent variables in each of your equations to determine what their influence is on your dependent variable.

To test the overall success of your model, you can determine the probability that you can predict the category of the dependent variable from the values of your independent variables. The

higher this probability is, the stronger the relationship is between the independent variables and your dependent variable. You can determine this probability iteratively using maximum likelihood estimation. If you multiply the logarithm of this probability by -2 , you will obtain a statistic that has an approximate chi-square distribution, with degrees of freedom equal to the number of parameters in your model. This is referred to as $-2LL$ (minus 2 log likelihood) and is commonly used to assess the fit of the model. Large values of $-2LL$ indicate that the observed model has *poor* fit. This statistic can also be used to provide a statistical test of the relationship between each independent variable and your dependent variable. The importance of each term in the model can be assessed by examining the increase in $-2LL$ when the term is dropped. This difference also has a chi-square distribution, and can be used as a statistical test of whether there is an independent relationship between each term and the dependent variable.

To performing a logistic regression in SPSS

- Choose **Analyze** → **Regression** → **Multinomial Logistic**.
- Move the categorical DV to the **Dependent** box.
- Move your categorical IVs to the **Factor(s)** box.
- Move your continuous independent variables to the **Covariate(s)** box.
- By default, SPSS does not include any interaction terms in your model. You will need to click the **Model** button and manually build your model if you want to include any interactions.
- When you are finished, you click the *Ok* button to tell SPSS to perform the analysis.

If your dependent variable only has two groups, you have the option of selecting **Analyze** → **Regression** → **Binary Logistic**. Though this performs the same basic analysis, this procedure is primarily designed to perform model building. It organizes the output in a less straightforward way and does not provide you with the likelihood ratio test for each of your predictors. You are therefore better off if you only use this selection if you are specifically interested in using the model-building procedures that it offers.

NOTE: The results from a binary logistic analysis in SPSS will actually produce coefficients that are opposite in sign when compared to the results of a multinomial logistic regression performed on exactly the same data. This is because the binary procedure chooses to predict the probability of choosing the category with the largest indicator variable, while the multinomial procedure chooses to predict the probability of choosing the category with the smallest indicator variable.

The **Multinomial Logistic** procedure will produce output with the following sections.

- **Case Processing Summary**. Describes the levels of the dependent variable and any categorical independent variables.
- **Model Fitting Information**. Tells you the $-2LL$ of both a null model containing only the intercept and the full model being tested. Recall that this statistic follows a chi-square distribution and that significant values indicate that there is a significant amount of variability in your DV that is *not* accounted for by your model.
- **Pseudo R-Square**. Provides a number of statistics that researchers have developed to represent the ability of a logistic regression model to account for variability in the dependent variable. Logistic regression does not have a true R-square statistic because

the amount of variance is partly determined by the distribution of the dependent variable. The more even the observations are distributed among the levels of the dependent variable, the greater the variance in the observations. This means that the R-square values for models that have different distributions are not directly comparable. However, these statistics can be useful for comparing the fit of different models predicting the same response variable. The most commonly reported pseudo R-square estimate is Nagelkerke's R-square, which is provided by SPSS in this section.

- **Likelihood Ratio Tests.** Provides the likelihood ratio tests for the IVs. The first column of the table contains the $-2LL$ (a measurement of model error having a chi-square distribution) of a model that does not include the factor listed in the row. The value in the first row (labeled **Intercept**) is actually the $-2LL$ for the full model. The second column is the difference between the $-2LL$ for the full model and the $-2LL$ for the model that excludes the factor listed in the row. This is a measure of the amount of variability that is accounted for by the factor. This difference parallels the Type III SS in a regression model, and follows a chi-square distribution with degrees of freedom equal to the number of parameters it takes to code the factor. The final column provides the p-value for the test of the null hypothesis that the amount of error in the model that excludes the factor is the same as the amount of error in the full model. A significant statistic indicates that the factor *does* account for a significant amount of the variability in the dependent variable that is not captured by other variables in the model.
- **Parameter Estimates.** Provides the specific coefficients of the logistic regression equations. You will have a number of equations equal to the number of levels in your dependent variable $- 1$. Each equation predicts the log odds of your observations being in the highest numbered level of your dependent variable compared to another level (which is listed in the leftmost column of the chart). Within each equation, you will see estimates of the standardized logistic regression coefficient for each variable in the model. These coefficients tell you the increase in the log odds when the variable increases by 1 (assuming everything else is held constant). The next column contains the standard errors of those coefficients. The Wald Statistic provides another statistic testing the significance of the individual coefficients, and is based on the relationship between the coefficient and its standard error. However, there is a flaw in this statistic such that large coefficients may have inappropriately large standard errors, so researchers typically prefer to use the likelihood ratio test to determine the importance of individual factors in the model. SPSS provides the odds ratio for the parameter under the column **Exp(B)**. The last two columns in the table provide the upper and lower bounds for a 95% confidence interval around the odds ratio.

RELIABILITY

Ideally, the measurements that we take with a scale would always replicate perfectly. However, in the real world there are a number of external random factors that can affect the way that respondents provide answers to a scale. A particular measurement taken with the scale is therefore composed of two factors: the theoretical "true score" of the scale and the variation caused by random factors. Reliability is a measure of how much of the variability in the observed scores actually represents variability in the underlying true score. Reliability ranges from 0 to 1. In psychology it is preferred to have scales with reliability greater than .7.

The reliability of a scale is heavily dependent on the number of items composing the scale. Even using items with poor internal consistency, you can get a reliable scale if your scale is long enough. For example, 10 items that have an average inter-item correlation of only .2 will produce a scale with a reliability of .714. However, the benefit of adding additional items decreases as the scale grows larger, and mostly disappears after 20 items. One consequence of this is that adding extra items to a scale will generally increase the scale's reliability, even if the new items are not particularly good. An item will have to significantly lower the average inter-item correlation for it to have a negative impact on reliability.

Reliability has specific implications for the utility of your scale. The most that responses to your scale can correlate with any other variable is equal to the square root of the scale's reliability. The variability in your measure will prevent anything higher. Therefore, the higher the reliability of your scale, the easier it is to obtain significant findings. This is probably what you should think about when you want to determine if your scale has a high enough reliability.

It should also be noted that low reliability does not call into question results obtained using a scale. Low reliability only hurts your chances of finding significant results. It cannot cause you to obtain false significance. If anything, finding significant results with an unreliable scale indicates that you have discovered a particularly strong effect, since it was able to overcome the hindrances of your unreliable scale. In this way, using a scale with low reliability is analogous to conducting an experiment with a small number of participants.

Calculating reliability from parallel measurements

One way to calculate reliability is to correlate the scores on parallel measurements of the scale. Two measurements are defined as parallel if they are distinct (are based on different data) but equivalent (such that you expect responses to the two measurements to have the same true score). The two measurements must be performed on the same (or matched) respondents so that the correlation can be performed. There are a number of different ways to measure reliability using parallel measurements. Below are several examples.

Test-Retest method. In this method, you have respondents complete the scale at two different points in time. The reliability of the scale can then be estimated by the correlation between the two scores. The accuracy of this method rests on the assumption that the participants are fundamentally the same (i.e., possess the same true score on your scale) during your two test periods. One common problem is that completing the scale the first time can change the way that respondents complete the scale the second time. If they remember any of their specific responses

from the first period, for example, it could artificially inflate the reliability estimate. When using this method, you should present evidence that this is not an issue.

Alternate Forms method. This method, also referred to as parallel forms, is basically the same as the Test-Retest method, but with the use of different versions of the scale during each session. The use of different versions reduces the likelihood that the first administration of the scale influences responses to the second. The reliability of the scale can then be estimated by the correlation between the two scores. When using alternate forms, you should show that the administration of the first scale did not affect responses to the second and that the two versions of your scale are essentially the same. The use of this method is generally preferred to the Test-Retest method.

Split-Halves method. One difficulty with both the Test-Retest and the Alternate Forms methods is that the scale responses must be collected at two different points in time. This requires more work and introduces the possibility that some natural event might change the actual true score between the two administrations of the scale. In the Split-Halves method you only have respondents fill out your scale one time. You then divide your scale items into two sections (such as the even-numbered items and the odd-numbered items) and calculate a score for each half. You then determine the correlation between these two scores. Unlike the other methods, this correlation does not estimate your scale's reliability. Instead, you get your estimate using the formula:

$$\hat{\rho} = \frac{2r}{1+r}$$

where $\hat{\rho}$ is the reliability estimate and r is the correlation that you obtain.

Note that if you split your scale in different ways, you will obtain different reliability estimates. Assuming that there are no confounding variables, all split-halves should be centered on the true reliability. In general it is best not to use a first half/second half split of the questionnaire since respondents may become tired as they work through the scale. This would mean that you would expect greater variability in the score from the second half than in the score from the first half. In this case, your two measurements are not actually parallel, making your reliability estimate invalid. A more acceptable method would be to divide your scale into sections of odd-numbered and even-numbered items.

Calculating reliability from internal consistency

The other way to calculate reliability is to use a measure of internal consistency. The most popular of these reliability estimates is *Cronbach's alpha*. Cronbach's alpha can be obtained using the equation:

$$\alpha = \frac{N\bar{r}}{1 + \bar{r}(N-1)},$$

where α is Cronbach's alpha, N is the number of items in the scale, and \bar{r} is the mean inter-item correlation. From the equation we can see that α increases both with increasing \bar{r} as well as with increasing N . Calculating Cronbach's alpha is the most commonly used procedure to estimate reliability. It is highly accurate and has the advantage of only requiring a single administration of the scale. The only real disadvantage is that it is difficult to calculate by hand, as it requires you to calculate the correlation between every single pair of items in your scale. This is rarely an issue, however, since SPSS will calculate it for you automatically.

To obtain the α of a set of items in SPSS:

- Choose **Analyze** → **Scale** → **Reliability analysis**.
- Move all of the items in the scale to the **Items** box.
- Click the **Statistics** button.
- Check the box next to **Scale if item deleted**.
- Click the **Continue** button.
- Click the **OK** button.

Note: Before performing this analysis, make sure all items are coded in the same direction. That is, for every item, larger values should consistently indicate either more of the construct or less of the construct.

The output from this analysis will include a single section titled **Reliability**. The reliability of your scale will actually appear at the bottom of the output next to the word **Alpha**. The top of this section contains information about the consistency of each item with the scale as a whole. You use this to determine whether there are any “bad items” in your scale (i.e., ones that are not representing the construct you are trying to measure). The column labeled **Corrected Item-Total Correlation** tells you the correlation between each item and the average of the other items in your scale. The column labeled **Alpha if Item Deleted** tells you what the reliability of your scale would be if you would delete the given item. You will generally want to remove any items where the reliability of the scale would increase if it were deleted, and you want to keep any items where the reliability of the scale would drop if it were deleted. If any of your items have a negative item-total score correlation it may mean that you forgot to reverse code the item.

Inter-rater reliability

A final type of reliability that is commonly assessed in psychological research is called “inter-rater reliability.” Inter-rater reliability is used when judges are asked to code some stimuli, and the analyst wants to know how much those judges agree. If the judges are making continuous ratings, the analyst can simply calculate a correlation between the judges’ responses. More commonly, judges are asked to make categorical decisions about stimuli. In this case, reliability is assessed via Cohen’s kappa.

To obtain Cohen's kappa in SPSS, you first must set up your data file in the appropriate manner. The codes from each judge should be represented as separate variables in the data set. For example, suppose a researcher asked participants to list their thoughts about a persuasive message. Each judge was given a spreadsheet with one thought per row. The two judges were then asked to code each thought as: 1 = neutral response to the message, 2 = positive response to the message, 3 = negative response to the message, or 4 = irrelevant thought. Once both judges

have rendered their codes, the analyst should create an SPSS data file with two columns, one for each judge's codes.

To obtain Cohen's kappa in SPSS

- Choose **Analyze** → **Descriptives** → **Crosstabs**.
- Place Judge A's responses in the **Row(s)** box.
- Place Judge B's responses in the **Column(s)** box.
- Click the **Statistics** button.
- Check the box next to **Kappa**.
- Click the **Continue** button.
- Click the **OK** button.

The output from this analysis will contain the following sections.

- **Case Processing Summary**. Reports the number observations on which you have ratings from both of your judges.
- **Crosstabulation**. This table lists all the reported values from each judge and the number of times each combination of codes was rendered. For example, assuming that each judge used all the codes in the thought-listing example (e.g., code values 1 – 4), the output would contain a cross-tabulation table like this:

Judge A * Judge B Crosstabulation

Count	Judge B				Total	
	1.00	2.00	3.00	4.00		
Judge A	1.00	5	1		6	
	2.00		5	1	6	
	3.00		1	7	8	
	4.00				7	
Total		5	7	8	7	27

The counts on the diagonal represent agreements. That is, these counts represent the number of times both Judges A and B coded a thought with a 1, 2, 3, or 4. The more agreements, the better the inter-rater reliability. Values not on the diagonal represent disagreements. In this example, we can see that there was one occasion when Judge A coded a thought in category 1 but Judge B coded that same thought in category 2.

- **Symmetric Measures**. The value of kappa can be found in this section at the intersection of the **Kappa** row and the **Value** column. This section also reports a p-value for the Kappa, but this is not typically used in reliability analysis.

Note that a kappa cannot be computed on a non-symmetric table. For instance, if Judge A had used codes 1 – 4, but Judge B never used code 1 at all, the table would not be symmetric. This is because there would be 4 rows for Judge A but only 3 columns for Judge B. Should you have this situation, you should first determine which values are not used by both judges. You then change each instance of these codes to some other value that is *not* the value chosen by the opposite judge. Since the original code was a mismatch, you can preserve the original amount of agreement by simply changing the value to a different mismatch. This way you can remove the

unbalanced code from your scheme while retaining the information from every observation. You can then use the kappa obtained from this revised data set as an accurate measure of the reliability of the original codes.

FACTOR ANALYSIS

Factor analysis is a collection of methods used to examine how underlying constructs influence the responses on a number of measured variables. There are basically two types of factor analysis: exploratory and confirmatory. Exploratory factor analysis (EFA) attempts to discover the nature of the constructs influencing a set of responses. Confirmatory factor analysis (CFA) tests whether a specified set of constructs is influencing responses in a predicted way. SPSS only has the capability to perform EFA. CFAs require a program with the ability to perform structural equation modeling, such as LISREL or AMOS.

The primary objectives of an EFA are to determine the number of factors influencing a set of measures and the strength of the relationship between each factor and each observed measure. To perform an EFA, you first identify a set of variables that you want to analyze. SPSS will then examine the correlation matrix between those variables to identify those that tend to vary together. Each of these groups will be associated with a factor (although it is possible that a single variable could be part of several groups and several factors). You will also receive a set of *factor loadings*, which tells you how strongly each variable is related to each factor. They also allow you to calculate *factor scores* for each participant by multiplying the response on each variable by the corresponding factor loading. Once you identify the construct underlying a factor, you can use the factor scores to tell you how much of that construct is possessed by each participant.

Some common uses of EFA are to:

- Identify the nature of the constructs underlying responses in a specific content area.
- Determine what sets of items "hang together" in a questionnaire.
- Demonstrate the dimensionality of a measurement scale. Researchers often wish to develop scales that respond to a single characteristic.
- Determine what features are most important when classifying a group of items.
- Generate "factor scores" representing values of the underlying constructs for use in other analyses.
- Create a set of uncorrelated factor scores from a set of highly collinear predictor variables.
- Use a small set of factor scores to represent the variable contained in a larger set of variables. This is often referred to as *data reduction*.

It is important to note that EFA does not produce any statistical tests. It therefore cannot ever provide concrete evidence that a particular structure exists in your data – it can only direct you to what patterns there may be. If you want to actually test whether a particular structure exists in your data you should use CFA, which does allow you to test whether your proposed structure is able to account for a significant amount of variability in your items.

EFA is strongly related to another procedure called principle components analysis (PCA). The two have basically the same purpose: to identify a set of underlying constructs that can account for the variability in a set of variables. However, PCA is based on a different statistical model, and produces slightly different results when compared to EFA. EFA tends to produce better results when you want to identify a set of latent factors that underlie the responses on a set of

measures, whereas PCA works better when you want to perform data reduction. Although SPSS says that it performs “factor analysis,” statistically it actually performs PCA. The differences are slight enough that you will generally not need to be concerned about them – you can use the results from a PCA for all of the same things that you would the results of an EFA. However, if you want to identify latent constructs, you should be aware that you might be able to get slightly better results if you used a statistical package that can actually perform EFA, such as SAS, AMOS, or LISREL.

Factor analyses require a substantial number of subjects to generate reliable results. As a general rule, the minimum sample size should be the larger of 100 or 5 times the number of items in your factor analysis. Though you can still conduct a factor analysis with fewer subjects, the results will not be very stable.

To perform an EFA in SPSS

- Choose **Analyze** → **Data Reduction** → **Factor**.
- Move the variables you want to include in your factor analysis to the **Variables** box.
- If you want to restrict the factor analysis to those cases that have a particular value on a variable, you can put that variable in the **Selection Variable** box and then click **Value** to tell SPSS which value you want the included cases to have.
- Click the **Extraction** button to indicate how many factors you want to extract from your items. The maximum number of factors you can extract is equal to the number of items in your analysis, although you will typically want to examine a much smaller number. There are several different ways to choose how many factors to examine. First, you may want to look for a specific number of factors for theoretical reasons. Second, you can choose to keep factors that have eigenvalues over 1. A factor with an eigenvalue of 1 is able to account for the amount of variability present in a single item, so factors that account for less variability than this will likely not be very meaningful. A final method is to create a *Scree Plot*, where you graph the amount of variability that each of the factors is able to account for in descending order. You then use all the factors that occur prior to the last major drop in the amount of variance accounted for. If you wish to use this method, you should run the factor analysis twice - once to generate the Scree plot, and a second time where you specify exactly how many factors you want to examine.
- Click the **Rotation** button to select a rotation method. Though you do not need to rotate your solution, using a rotation typically provides you with more interpretable factors by locating solutions with more extreme factor loadings. There are two broad classes of rotations: orthogonal and oblique. If you choose an orthogonal rotation, then your resulting factors will all be uncorrelated with each other. If you choose an oblique rotation, you allow your factors to be correlated. Which you should choose depends on your purpose for performing the factor analysis, as well as your beliefs about the constructs that underlie responses to your items. If you think that the underlying constructs are independent, or if you are specifically trying to get a set of uncorrelated factor scores, then you should clearly choose an orthogonal rotation. If you think that the underlying constructs may be correlated, then you should choose an oblique rotation. *Varimax* is the most popular orthogonal rotation, whereas *Direct Oblimin* is the most popular oblique rotation. If you decide to perform a rotation on your solution, you usually ignore the parts of the output that deal with the initial (unrotated) solution since

the rotated solution will generally provide more interpretable results. If you want to use direct oblimin rotation, you will also need to specify the parameter *delta*. This parameter influences the extent that your final factors will be correlated. Negative values lead to lower correlations whereas positive values lead to higher correlations. You should not choose a value over .8 or else the high correlations will make it very difficult to differentiate the factors.

- If you want SPSS to save the factor scores as variables in your data set, then you can click the **Scores** button and check the box next to **Save as variables**.
- Click the **Ok** button when you are ready for SPSS to perform the analysis.

The output from a factor analysis will vary depending on the type of rotation you chose. Both orthogonal and oblique rotations will contain the following sections.

- **Communalities**. The communality of a given item is the proportion of its variance that can be accounted for by your factors. In the first column you'll see that the communality for the initial extraction is always 1. This is because the full set of factors is specifically designed to account for the variability in the full set of items. The second column provides the communalities of the final set of factors that you decided to extract.
- **Total Variance Explained**. Provides you with the eigenvalues and the amount of variance explained by each factor in both the initial and the rotated solutions. If you requested a Scree plot, this information will be presented in a graph following the table.
- **Component Matrix**. Presents the factor loadings for the initial solution. Factor loadings can be interpreted as standardized regression coefficients, regressing the factor on the measures. Factor loadings less than .3 are considered weak, loadings between .3 and .6 are considered moderate, and loadings greater than .6 are considered to be large.

Factor analyses using an orthogonal rotation will include the following section.

- **Rotated Component Matrix**. Provides the factor loadings for the orthogonal rotation. The rotated factor loadings can be interpreted in the same way as the unrotated factor loadings.
- **Component Transformation Matrix**. Provides the correlations between the factors in the original and in the rotated solutions.

Factor analyses using an oblique rotation will include the following sections.

- **Pattern Matrix**. Provides the factor loadings for the oblique rotation. The rotated factor loadings can be interpreted in the same way as the unrotated factor loadings.
- **Structure Matrix**. Holds the correlations between the factors and each of the items. This is not going to look the same as the pattern matrix because the factors themselves can be correlated. This means that an item can have a factor loading of zero for one factor but still be correlated with the factor, simply because it loads on other factors that are correlated with the first factor.
- **Component Correlation Matrix**. Provides you with the correlations among your rotated factors.

After you obtain the factor loadings, you will want to come up with a theoretical interpretation of each of your factors. You define a factor by considering the possible constructs that could be responsible for the observed pattern of positive and negative loadings. You should examine the

items that have the largest loadings and consider what they have in common. To ease interpretation, you have the option of multiplying all of the loadings for a given factor by -1. This essentially reverses the scale of the factor, allowing you, for example, to turn an "unfriendliness" factor into a "friendliness" factor.

VECTORS AND LOOPS

Vectors and loops are two tools drawn from computer programming that can be very useful when manipulating data. Their primary use is to perform a large number of similar computations using a relatively small program. Some of the more complicated types of data manipulation can only reasonably be done using vectors and loops.

A vector is a set of variables that are linked together because they represent similar things. The purpose of the vector is to provide a single name that can be used to access any of the entire set of variables. A loop is used to tell the computer to perform a set of procedures a specified number of times. Often times we need to perform the same transformation on a large number of variables. By using a loop, we only need to define the transformation once, and can then tell the computer to do the same thing to all the variables using a loop.

If you have computer-programming experience then you have likely come across these ideas before. However, what SPSS calls a “vector” is typically referred to as an “array” in most programming languages. If you are familiar with arrays and loops from a computer-programming course, you are a step ahead. Vectors and loops are used in data manipulation in more or less the same way that arrays and loops are used in standard computer programming.

Vectors

Vectors can only be defined and used in syntax. Before you can use a vector you first need to define it. You must specify the name of the vector and list what variables are associated with it. Variables referenced by a vector are called “elements” of that vector. You declare a vector using the following syntax.

```
vector Vname = varX1 to varX2.
```

If the variables in the vector have not already been declared, you can do so as part of the vector statement. For more information on this, see page 904 of the *SPSS Base Syntax Reference Guide*. The following are all acceptable vector declarations.

```
vector V = v1 to v8.
vector Myvector = entry01 to entry64.
vector Grade = gradel1 to gradel2.
vector Income = in1992 to in2000.
```

The vector is given the name *Vname* and is used to reference a set of variables defined by the variable list. The elements in the vector must be declared using the syntax *first variable to last variable*. You cannot list them out individually. This means that the variables to be included in a vector must all be grouped together in your data set.

Vectors can be used in transformation statements just like variables. However, the vector itself isn't able to hold values. Instead, the vector acts as a mediator between your statement and the variables it references. The variables included in a vector are placed in a specific order, determined by the declaration statement. So if you give SPSS a vector and an order number (referred to as the *index*), it knows what specific element you want to access. You do not need to

know what the exact name of the variable is - you just need to know its location in the vector. References to items within a vector are typically made using the format

```
vname (index)
```

where *vname* is the name of the vector, and *index* is the numerical position of the desired element. Using this format, you can use a vector to reference a variable in any place that you would normally insert a variable name. For example, all of the following would be valid SPSS statements, assuming that we had defined the four variables above.

```
compute V(4) = 6.
if (Myvector(30)='house') correct = correct + 1.
compute sum1 = Grade(1) + Grade(2) + Grade(3).
compute change = Income(9) - Income(1).
```

Note that the index used by a vector only takes into account the position of elements in the vector - not the names of the variables. To reference the variable *in1993* from in the *Income* vector above, you would use the phrase `income(2)`, not `income(1993)`.

Using vectors this way doesn't provide us with much of an advantage - we are not really saving ourselves any effort by referring to a particular variable as *Myvector(1)* instead of *entry01*. The advantage comes in with the fact that the index of the vector itself can be a variable. In this case, the element that the vector will reference will depend on the value of the index variable. So the exact variable that is changed by the statement

```
compute Grade(t) = Grade(t) + 1.
```

depends on the value of *t* when this statement is executed. If *t* has the value of 1, then the variable *grade1* will be incremented by 1. If *t* has a value of 8, then the variable *grade8* will be incremented by 1. This means that the same statement can be used to perform many different things, simply depending what value you assign to *t*. This allows you to use vectors to write “generic” sections of code, where you control exactly what the code does by assigning different values to the index variables.

Loops

Vectors are most useful when they are combined with loops. A loop is a statement that lets you tell the computer to perform a set of commands a specified number of times. In SPSS you can tell the computer to perform a loop by using the following code:

```
loop loop_variable = lower_limit to upper_limit.
--commands to be repeated appear here--
end loop.
```

When SPSS encounters a loop statement, what it does first is set the value of the loop variable to be equal to the lower limit. It then performs all of the commands inside the loop until it reaches the `end loop` statement. At that point the computer adds 1 to the loop variable, and then compares it to the upper limit. If the new value of the loop variable is less than or equal to the upper limit, it goes back to the beginning of the loop and goes through all of the commands

again. If the new value is greater than the upper limit, the computer then moves to the statement after the `end loop` statement. Basically, this means that the computer performs the statements inside the loop a total number of times equal to (upper limit – lower limit + 1).

The following is an example of an SPSS program that uses a loop to calculate a sum:

```
compute x = 0.
loop #t = 4 to 8.
+   compute x = x + #t.
end loop.
```

The first line simply initializes the variable count to the value of zero. The second line defines the conditions of the loop. The loop variable is named *t*, and starts with a value of 4. The loop cycles until the value of *t* is greater than 8. This causes the program to perform a total of 5 cycles. During each cycle the current value of *t* is added to *x*. At the end of this set of statements, the variable *x* would have the value of $4 + 5 + 6 + 7 + 8 = 30$.

In this example, the loop variable is denoted as a “scratch variable” because its first letter is a number sign (#). When something is denoted as a scratch variable in SPSS it is not saved in the final data set. Typically we are not interested in storing the values of our loop variables, so it is common practice to denote them as scratch variables. For more information on scratch variables see page 32 of the *SPSS Base Syntax Reference Guide*.

You will also notice the plus sign (+) placed before the compute statement in line 3. SPSS needs you to start all new commands in the first column of each line. Here we wish to indent the command to indicate that it is part of the loop. We therefore put the plus symbol in the first column which tells SPSS that the actual command starts later on the line.

Just in case you were wondering, the first statement setting $x = 0$ is actually necessary for the sum to be calculated. Most programming languages, including SPSS syntax, start variables with missing values. Adding anything to a missing value produces a missing value, so we must explicitly start the variable count at zero to be able to obtain the sum.

The Power of Combining Vectors and Loops

Though you can work with vectors and loops alone, they were truly designed to be used together. A combination of vectors and loops can save you incredible amounts of time when performing certain types of repetitive transformations. Consider the characteristics of vectors and loops. A vector lets you reference a set of related variables using a single name and an index. The index can be a variable or a mathematical expression involving one or more variables. A loop repeatedly performs a set of commands, incrementing a loop variable after each cycle. What would happen if a statement inside of a loop referenced a vector using the loop variable as the index? During each cycle, the loop variable increases by 1. So during each cycle, the vector would refer to a different variable. If you correctly design the upper and lower limits of your loop, you could use a loop to perform a transformation on every element of a vector.

For an example, let's say that you conducted a reaction-time study where research participants observed strings of letters on the screen and judged whether they composed a real word or not. In your study, you had a total of 200 trials in several experimental conditions. You want to analyze

your data with an ANOVA to see if the reaction time varies by condition, but you find that the data has a right skew (which is common). To use ANOVA, you will need to transform the data so that it has a normal distribution, which involves taking the logarithm of the response time on each trial. In terms of your data set, what you need is a set of 200 new variables whose values are equal to the logarithms of the 200 response time variables. Without using vectors or loops, you would need to write 200 individual transformation statements to create each log variable from the corresponding response time variable. Using vectors and loops, however, we can do the same work with the following simple program. The program assumes that the original response time variables are *rt001* to *rt200*, and the desired log variables will be *lrt001* to *lrt200*.

```
vector Rvector = rt001 to rt200.
vector Lvector = lrt001 to lrt200.
loop #item = 1 to 200.
+   compute Lvector(#item) = log(Rvector(#item)).
end loop.
```

The first two statements set up a pair of vectors, one to represent the original response time variables and one to represent the transformed variables. The third statement creates a loop with 200 cycles. Each cycle of the loop corresponds to a trial in the experiment. The fourth line actually performs the desired transformation. During each cycle it takes one variable from `Lvector` and sets it equal to the log of the corresponding variable in `Rvector`. The fifth line simply ends the loop. By the time this program completes, it will have created 200 new variables holding the log values that you desire.

In addition to greatly reducing the number of programming lines, there are other advantages to performing transformations using vectors and loops. If you need to make a change to the transformation you only need to change a single statement. If you write separate transformations for each variable, you must change every single statement anytime you want to change the specifics of the transformation. It is also much easier to read programs that use loops than programs with large numbers of transformation statements. The loops naturally group together transformations that are all of the same type, whereas with a list you must examine each individual transformation to find out what it does.