Applications of Intermediate/Advanced Statistics in Institutional Research

Edited by
Mary Ann Coughlin

THE ASSOCIATION FOR INSTITUTIONAL RESEARCH
Number Sixteen
Resources in Institional Research

222 Stone Building
Florida State University
Tallahassee, FL 32306-4462

Printed in the United States

ISBN 882393-12-0

Dedication

The authors of this monograph dedicate this work to the memory of Julia Duckwall. Julia touched the lives of so many Institutional Research professionals with her spirit and dedication to the profession. We hope that this monograph serves as a valued reference for many and that the work and spirit of Julia will live on.

Table of Contents

Introduction

In March of 1999, the Association for Institutional Research offered the first Applied Statistics Institute. The Professional Development Services Committee in conjunction with the Publications Committee undertook the development of the current *Resources in Institutional Research* monograph. The goal of this document is to provide a resource for institutional research professionals concerning the application of intermediate/advanced statistics in institutional research settings, as well as provide a resource document to participants attending the Applied Statistics Institute.

As a result, the curriculum of the Applied Statistics Institute has served as the basis for the content of this monograph. The Institute offers five specialized modules. Each module provides a theoretical context with practical applications, exercises, and interpretive and presentation techniques for each statistical approach. The five modules focus on: non-parametric statistics, regression analysis, analysis of variance, identifying and analyzing group difference, multilevel models, and multivariate statistics. As a result, each chapter is authored by the faculty member who has described these applications and techniques. Additional data sets and exercises will be made available on the AIR Web site www.airweb.org.

The focus of this monograph is not to cover each statistical area in depth; rather it is to describe the theory and application of these procedures to institutional research settings. As a result, the reader should be familiar with basic statistical principles and applications. In addition, the reader may need to refer to supplemental readings provided within each chapter to more fully understand each statistical application.

Similar to the learning objectives of the Applied Statistics Institute, the goal of this monograph is to educate the reader about: uses of non-parametric statistics for common assessment activities; applications of regression techniques to higher education problems and issues; uses of ANOVA for rating scale data, student performance data, and other IR data; applications of techniques for identifying groups and determining how groups differ; uses of advanced statistics to provide evidence of institutional effectiveness; and applications of multilevel modeling techniques to common institutional research questions.

ENJOY and consider joining us for an upcoming Applied Statistics Institute.

Mary Ann Coughlin
Editor

Chapter 1
Nonparametric Statistics: Applications in Institutional Research

Richard Howard
Gerald McLaughlin
Josetta McLaughlin

The statistical tests and procedures outlined in other chapters of this book in general are known as "parametric tests." Those statistical procedures require the researcher to assume that the population from which data are collected reflect a normal distribution and, assuming that the sample is representative of the population, also reflect the properties of a normal distribution. While the actual distribution of the sample may not be exactly normal, it is considered "close enough" in most cases, and the problem under study is modeled using the assumptions and probabilities that define the normal distribution. In this way, the use of *Parametric Statistics results in exact solutions to approximate problems* (Conover, 1971). Computationally, parametric tests require: (1) sample sizes greater than 30 observations; (2) data that reflect the properties of interval or ratio measurement scales; and (3) specific data about each observation.

While appropriate for many research projects, parametric statistics do not serve the needs of researchers whose data sets fail to meet the criteria noted above or whose data sets are small due to the nature of the project. During the 1930s, statisticians proposed alternative procedures that did not rely on the assumptions required to use parametric statistics. The resultant statistical tests, known as nonparametric tests, are not dependent on the normal distribution to define desired probabilities but use other distributions or close approximations. These tests allow the researcher to model the problem under study. In many cases, they are easier to apply in that less computational work is required. As such, *Nonparametric Statistics results in approximate solutions to exact problems* (Conover, 1971). Computationally, these tests: (1) are not dependent on large numbers of observations; (2) use data that reflect in most cases the properties of nominal or ordinal measurement scales; and (3) are frequently used to analyze summarized or categorical count data.

The choice of whether to use a parametric test or nonparametric test is dictated by the characteristics of the data as described above. Be aware, however, that nonparametric tests are not in general as sensitive as parametric tests. In other words, parametric tests are more likely than their nonparametric counterparts to detect a statistically significant difference between two or more treatments or a significant relationship between two variables. When

faced with a situation where the data will allow you to choose between the use of a parametric test and a nonparametric test, the parametric test is the recommended option (Gravetter & Wallnau, 2004; Zar, 1984).

Institutional Research Context

Often the statistical problems that face institutional researchers include situations involving large numbers of observations such as the student population or the institution's faculty and staff. In these cases, the use of parametric approaches to studying the problems are usually appropriate as the underlying distributions are typically "close enough" to normal to provide reliable information. However, it is also often the case that the data reflect one or more of the following characteristics making the use of parametric analyses inappropriate or impossible:

- small sample size;
- data summarized into categories;
- a non-normal or unknown distribution; and/or
- nominal or ordinal measurement scales.

Institutional researchers are often faced with situations when the data they are working with originate in reports (paper and Web-based) in which the data are summarized in categories, such as disciplines, students by rank, or by institutions. In other cases, the unit of analysis might be the department, college, or program in which comparison information between six to ten comparators is the intent of the analysis. The normal distribution does not model the data, especially when there are not enough observations to invoke the assumptions of normality and parametric tests are not appropriate. Nonparametric tests were designed specifically to address these situations.

In this chapter, a number of nonparametric tests are presented with examples reflecting "typical" questions that might be asked of an institutional research office. The primary and traditional non-parametric tests included are those that have the following characteristics: (1) standard procedures exist to compute them and (2) they are included in the SPSS procedures. Many of these tests have large sample equations, but we do not present those formulae in this chapter. They can be found in basic nonparametric texts such as The Handbook of Parametric and Nonparametric Statistical Procedures, by D. J. Sheskin, 1997. In addition, we discuss some fairly new and advanced techniques. Some, such as *log-linear analysis*, are statistical tests. Others, such as *Bootstrapping*, can lead to statistical statements. Finally, some of the techniques, such as *Data Mining,* do not tend to make probabilistic statements but are more an extension of *Exploratory Data Analysis* techniques. Individuals interested in learning more about these tools are referred to the references at the end of the chapter.

Chapter Organization

The format of this chapter is somewhat different from the others in this book. In general, the other authors dealt with a specific family of tests, i.e. Analysis of Variance or Regression. In contrast, we present a number of tests that have three fundamental purposes – tests of "location", tests for "goodness of fit", and tests of "association". The only common parameter is that the tests do not rely on assumptions associated with **normal or other distributions**. In most introductory statistics and social science research texts, nonparametric statistical tests are discussed in terms of this assumption, and only the most common of the nonparametric tests are presented (Chi Square, Mann-Whiteny U, Sperman Rho, etc.). There are a number of situations where less common nonparametric tests can be used to provide the statistical evidence to support an institutional position, the evaluation of a policy, or the effects of a process or procedure. Obviously, the scope of this book does not allow us to present all nonparametric tests that might be appropriate for examining institutional data. Nevertheless, our intent is to provide an overview of selected tests with examples of how they can be used to answer typical questions posed to an institutional research office.

The order of the chapter is as follows: First we present a basic methodology for testing the assumption that the data to be analyzed reflects a normal distribution. Next we present a series of nonparametric tests appropriate for use with those cases where a normal distribution is not assumed, the scale is not interval or ratio, or other reasons exist that support using a nonparametric test. For each test, we indicate the purpose of the test, assumptions about the data, the hypothesis to be tested, an example of using the test in an institutional research context, and the SPSS procedure and output. The specific tests are presented and organized according to the purpose of the test and the number of samples.

The tests described in the chapter are summarized in Tables 1, 2, and 3 based on purpose, scale, and the number of samples. For each of the nonparametric tests presented in the tables, a data set has been developed and can be accessed through the Association for Institutional Research Web site (http://airweb.org). The data sets specific to particular tests are identified in the discussion of each test. The intent is that the reader should be able to access a particular data set and run the test as described in the chapter. This will allow the researcher to practice setting up the SPSS procedure and to then compare the outcome with that presented in the chapter. If the same output is obtained as shown in the chapter, then the researcher will be ready to analyze the data set of concern. Finally, we discuss some advanced methodologies and concerns.

Table 1
Nonparametric Tests of Location

No. of Samples / Scale	One Sample	Two Samples Independent	Two Samples Related	Three or more Samples Independent	Three or more Samples Related
Nominal	Binomial Test Runs Test		McNemar Test		Cochran Q
Ordinal	Sign Test	Median Test Mann-Whitney U Test	Sign Test for Two Dependent Samples	Median Test Kruskal-Wallis ANOVA by Ranks	Friedman Two-way ANOVA
Interval	Wilcoxon Signed-ranks Test		Wilcoxon Matched Pairs Signed-ranks Test		
Parametric Equivalent	One Sample t-test	Two Sample t-test	Paired t-test	ANOVA	Within Subjects ANOVA

Table 2
Nonparametric Analysis for Goodness of Fit*

No. of Samples / Scale	One Sample	Two Samples
Nominal	One Sample Chi Square	Chi Square Test of Independence
Ordinal	One Sample Kolmogorov-Smirnov	Two Sample Kolmogorov-Smirnov
*no generally comparable Parametric techniques		

Table 3
Nonparametric Analysis for Association

No. of Variables / Scale	Two Variables	Three or more Variables
Nominal	Phi Coefficient (2x2) Point Biserial (2xLinear) Chi Square Test of Independence	Log-Linear (not include)
Ordinal	Spearman Rho	Kendall's Coefficient of Concordance W
Parametric Equivalent	Pearson Correlation	Eta Squared

Looking at the Data: Determining When to Use Nonparametric Statistics

Properties of Nonparametric Statistics

As indicated above, nonparametric statistical tools are not based on the properties of the normal distribution. Because they do not require that assumptions be made about the normality of the sampled population, the term distribution-free test is sometimes applied to these statistical tools (Zar, 1984, p. 138).

Nonparametric statistics are those which have one or more of the following properties:

- The data are count data that enumerate the number of observations having some characteristic or belonging to a specific group.
- The data are measured and/or analyzed using a nominal scale or ordinal scale.
- The inference does not concern a parameter in the population.
- The probability distribution of the statistic on which the analysis is based is not dependent upon specific information or assumptions about the population from which the sample(s) is drawn, but only on general assumptions such as being continuous and/or symmetric (See Sheskin, 1997; Zar, 1984; Gibbons, 1971).

Measurement Scales

To know when it is appropriate to use a nonparametric test, the researcher must understand the level of measurement used to measure the characteristic of interest. As a quick review, we briefly define each of the four measurement scales commonly employed — nominal, ordinal, interval, and ratio. For a more detailed discussion of the four measurement scales, refer to any introductory statistics or social science research text (Gravetter & Wallnau, 2004; Hinkle, Wiersma, & Jurs, 1998; Gay & Airasian, 2003).

- Nominal: When the variable is classified on the basis of some quality rather than on a numerical basis, the level of measurement is nominal. An example would be a student's major. When using nominal data, the researcher generally counts the number of observations in each category. This level of measurement is sometimes referred to as categorical.
- Ordinal: When data reflect relative differences rather than quantitative differences and can be ranked, the level of measurement

is ordinal. The researcher generally ranks the data along some characteristic, typically from highest to lowest. An example would be student academic level, e.g., senior>junior>sophomore>first year students.

- Interval: When the elements can be differentiated and ordered and the arithmetic difference between elements is meaningful, the level of measurement is interval. The data possess a constant interval size but do not possess a true zero. An example would be student grades. (Even though this scale has a zero, it is not likely a "true" zero.)

- Ratio: When there is an interval scale with a fixed origin and the basic scale indicates proportionality, the level of measurement is ratio. In other words, an absolute zero point exists in the scale. An example would be tuition revenue.

Generally, nonparametric statistics require only that the data to be analyzed are nominal or ordinal. It is useful to remember that **the measurement scales are cumulative in that each scale involves the characteristics of the former scale plus another property.** As such, interval and ratio data can be restructured, i.e., recoded, as either nominal or ordinal data, and nonparametric statistical tests can then appropriately be used as the analysis tools. This is typically done when the number of observations is small or the distribution of the data is not normal.

Hypothesis Testing

As noted in other chapters of this book, a hypothesis is a statement, often based on some theory that is made to explain certain observations requiring further investigation. Setting up and testing hypotheses is a critical step in conducting credible statistical procedures. In the testing of a hypothesis, the first step is to determine the correct directional representation of the hypothesis. Typically one has the option of doing either a non-directional hypothesis ($X = Y$) or a directional hypothesis ($X>Y$). The safest option is the non-directional hypothesis. However, because the assigned risk is in both tails of the probability distribution of the statistic in the non-directional while only in one tail for the directional hypothesis, the directional hypothesis is more likely to be rejected. In other words, the statistic required for rejection of a non-directional hypothesis needs to be larger than the statistic required for rejection of a directional hypothesis at a given risk level.

After the decision on the hypothesis has been made, the next step is to select an alpha level that reflects the willingness to reject the null hypothesis when it is in fact true (Type I or α risk). There are then two basic ways to look at the appropriateness of rejecting or not rejecting the null hypothesis. They

differ only by a mathematical process. The first way is to use the rejection level (either $\alpha = p = .0x$ or $\alpha/2 = p = .0x$) to identify the associated critical value of the statistic. The second way is to take the computed statistic and determine the likelihood that such a statistic could occur by chance. This likelihood – or probability – is then compared to the alpha risk for the hypothesis, and a decision is made about whether to reject the null hypothesis. The traditional approach has been to compute a statistic and compare it to the value of the statistic required to make the determination about rejecting or not rejecting the null hypothesis. In fact, the output from SPSS makes the comparison of the probability of the computed statistic to the assigned alpha risk the most direct manner for making a decision about the null hypothesis.

Tests for Normality

As noted, choice of nonparametric statistics requires that the distribution of the data be evaluated. Evaluation of skewness and kurtosis can help the researcher determine whether the data reflect a normal or non-normal distribution. *Skewness* is a measure of asymmetry of the distribution of numbers. A normal distribution is symmetrical. There are multiple procedures for evaluation of skewness. For example, the formula for the mean of the cubed Z scores is sometimes used to calculate skewness. If the calculated value is "0", i.e., zero, the distribution of the set of numbers is symmetrical. If the value is less than zero, the long tail of the distribution is to the left of the distribution, and the mean is less than the median. If the value is greater than zero, the long tail is to the right and the mean is greater than the median (Mertler & Vannatta, 2002).

As with skewness, there are multiple procedures available for evaluating kurtosis. *Kurtosis* is a measure of the presence of extreme values in the distribution. If the distribution is relatively peaked in the middle, kurtosis will be greater than zero. If the distribution is rather flat, the kurtosis will be less than zero. Normal distributions have a kurtosis of zero (Mertler & Vannatta, 2002). The tests for skewness and kurtosis are very sensitive. Often a single number or a few numbers in a large distribution will result in a statistically significant skewness or kurtosis statistic. Inspecting a frequency distribution will reveal the outliers.

Example: A group of entering first year students was administered pre-tests upon registering for a basic linear algebra course. A post-test was administered following completion of the course. Researchers want to answer the following question: *For a group of entering first year students, is the distribution of the improvement scores on the math achievement test normal?*

In SPSS, the Skewness and the Kurtosis statistics are calculated using the Explore procedure. The Standard Error (SE) for each of these statistics is also calculated. To determine if a distribution of numbers is statistically different from a normal distribution, the Skewness and Kurtosis statistics can be converted to Z scores by dividing the statistic by its standard error.

The scores can then be compared to 1.96 (a = .05) or 2.56 (a = .01). If the Z value is less than these figures or critical values, it can be assumed that the distribution of the set of scores is normally distributed.

Conclusion: The statistical measures for skewness and kurtosis are not significantly different from what would be expected from a normal distribution.

The findings indicate that the distribution is normal or very close to normal. Dividing the statistic for skewness (.211) by its standard error (.121) yields 1.74, which is less than the critical value 1.96.

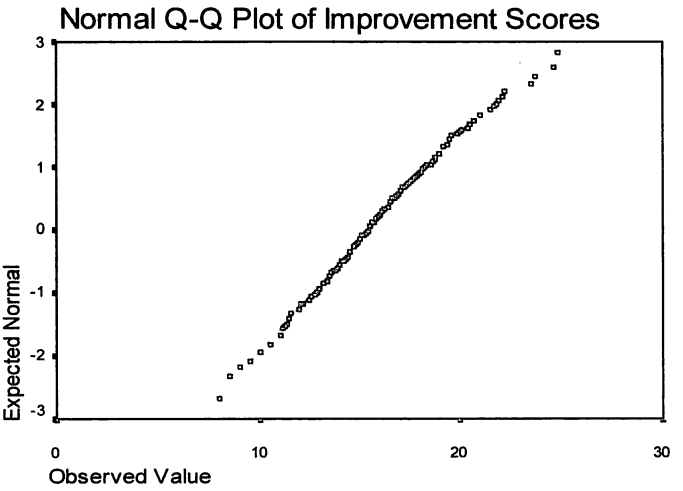
The Stem and Leaf output and Q-Q plot confirm this conclusion. The "stem and leaf" distribution is normal looking. In the Stem and Leaf plot, the "&" represents a fractional leaf. The trimmed mean is the mean of the observations between the 5th and 95th percentile and tends to be less sensitive to extreme values. In the Q-Q plot shown below, the observations are very near the line reflecting a normal distribution.

SPSS Output

		Statistic	Standard. Error
Improvement Scores	Mean	15.50	.140
	95% Confidence Interval for Mean	15.22	
		15.77	
	5% Trimmed Mean	15.46	
	Medium	15.50	
	Variance	7.958	
	Std. Deviation	2.821	
	Minimum	8	
	Maximum	25	
	Range	17	
	Interquartile Range	3.53	
	Skewness	**.211**	**.121**
	Kurtosis	**.389**	**.242**

Stem and Leaf Output for Math Improvement Scores

Frequency	Stem and Leaf
2.00	Extremes (=<8.0)
3.00	8. 5
3.00	9. 5&
6.00	10. 005
23.00	11. 000045555&&
28.00	12. 0000025555689&
43.00	13. 000000222455555556789
65.00	14. 00000000234455555555557789999&
62.00	15. 0000000123445555555555778889&
55.00	16. 000000002244445555567899&
39.00	17. 00000002334556678&
26.00	18. 00002255667&
27.00	19. 000000245556&
8.00	20. 15&
9.00	21. 00&&
3.00	22. 2&
4.00	Extremes (>=23.5)



SPSS Procedure:
Use data set: **tests of normality**

1. Descriptive Statistics
2. Explore
3. Move *Math Placement Test* from the "Variable List" window to the "Dependent List" window
4. Under "Display" click "Both"
5. Click "Plots"
6. Under "Boxplots" click "none"
7. Under "Descriptive" check "stem-and-leaf"
8. Check "Normality plots with tests"
9. Click "continue"
10. Click "OK"

Nonparametric Tests: Descriptions and Examples

Tests of Location: One Sample

Binomial Test: Nominal

A test that requires the presence of a dichotomous measure is the Binomial Test. This test is based on the characteristics of the Binomial Distribution. The assumption underpinning this distribution is that if an event is occurring at random from a distribution where an event has the likelihood of (p), then for N observations the event is expected to occur (N*p) times with a

variance of N*p*(1-p). Using this distribution in its known quantities, we can test to determine the likelihood an event is occurring from a population that has a hypothesized probability of (p). This is a one sample test of location for beliefs about the size of the probability of one category occurring in the population.

Assumptions:

1. The outcome can be categorized as a dichotomous measure where one of the categories occurs in the sample with probability p.
2. There is a hypothesized proportion of the outcomes (p_o) that will be in the specific categories.

Hypotheses:

Two-tailed: $H_o: p = p_o$, $H_1: p \neq p_o$,

One-tailed: $H_o: p \geq p_o$,$H_1: p < p_o$, or

$H_o: p \leq p_o$, $H_1: p > p_o$

Procedure:

1. State the null and alternative hypotheses.
2. Select a level of significance α.
3. Compute the proportion in the sample (p).
4. Reject the null hypothesis if the proportion is sufficiently small or large compared to the proportion in the hypothesis (p_o).

Note: When there are more than about twenty observations, an approximation using the normal distribution works well.

Example: It is observed that six of the students in an English literature class of twenty students are engineering majors. It is known that 60% of the students in the university are in Engineering. Researchers want to answer the following question: *Does this English class have a representative number of engineering majors?* ($p_o = .6$, $p = 6/20 = .3$, $n = 20$)

Hypothesis (Two Tailed): $H_o: p = p_o$, $H_1: p \neq p_o$ $\alpha = .05$

SPSS Output

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (1-tailed)
Major	Group 1	Engineers	6	.3	.6	**.006(a)**
	Group 2	Other	14	.7		
	Total		20	1.0		

a Alternative hypothesis states that the proportion of cases in the first group < .6.

SPSS Procedure:
Use data set: **Binomial Test**

1. Nonparametric Tests
2. Binomial Test
3. Move *Majors* from "Variables list" window to "'Dependent Variables List" window
4. Under "Defined Dichotomy" click "Get from data"
5. "Test Proportion" enter .60
6. Click "OK"

Conclusion: Reject the null hypothesis (H_o: $p = p_o$) because $p = .012 < .025$.

The hypothesis was stated as a two-tailed, nondirectional test (H_o: $p = p_o$, H_1: $p \neq p_o$). The SPSS Output is for a one-tailed, directional test (H_o: $p \geq p_o$, H_1: $p < p_o$). Therefore, the "Exact Sig." ($p = .006$) must be adjusted to reflect ($p = .012$). Results of the analysis thus suggest that the null hypothesis should be rejected. The researcher can conclude that the class does not have a representative number of engineering majors.

Runs Test for Random Sequence: Nominal

A Runs Test is a special case of the binomial test that examines non-randomness when there is a sequence of binary events, i.e., to determine if the sequence of events is random. For example, it may be desirable to see if the time to complete a learning task is related to a student's major. In this case, the score is 1 if the student is a business major and 2 if the student is any other type of major. This is similar in concept to looking at the normality of the distribution in that it looks at one of the more basic assumptions made when we do research. The assumption concerns randomness and whether the series of events is coming from a random sequence that would produce a binary sequence where the two events are occurring in a random sequence. If the two categories that are feasible are scored 1 and 2, the sequence of 1's and 2's should alternate a certain number of times by chance if the sequence is random.

A "run" is defined as a sequence of like items that are followed or preceded by a different item or no item at all. If there is a much smaller number or larger number of runs than one would expect, then the likelihood is that the sequence is not random. The researcher must thus examine the sequence in which the events are observed. The sequence can be a natural event, such as time, or it can be an ordered set of events such as the points on a trend line or regression equation. For large samples, one would expect 2p*q*n runs where p is the portion in one of the categories, q is the portion in the other ($q = 1 - p$), and n is the number of observations. A similar methodology also exists for looking at the length of the longest run but is not shown here.

Assumption:

There is a sequence of events that results in a binomial set of categories.

Hypotheses:

Two-tailed: H_o: $R = R_o$, H_1: $R \neq R_o$,

One-tailed: H_o: $R \geq R_o$, H1: $R < R_o$, or H_o: $R \leq R_o$, H_1: $R > R_o$

Procedure:

1. State the null and alternative hypotheses.
2. Select a level of significance α.
3. Count the number of times there is a run in the sequence of observations. (If the observations are occurring in a random sequence the number of runs will be about n*p*q.)
4. Determine whether there is a sufficiently small number of runs or a sufficiently large number of runs to reject the null hypothesis.

Example: When the time to complete an exam was identified, the pattern of majors (M) and non-majors (N) was M,M,M, N,N, M,M,M, N,N,N,N,N, M. The sequence of results has five runs (R = 5) and fourteen observations with seven majors and seven non-majors. Researchers want to answer the following question: *Does the pattern indicate that there is a difference in the time students take to complete the exam based on what their majors are?*

Hypotheses (Two Tailed): H_o: $R = R_o$, H_1: $R \neq R_o$, $\alpha = .05$

Conclusion: Fail to reject the null hypothesis (H_o: $R = R_o$) given that $p = .155 > .05$.

Results from SPSS give a probability of (p = .155) for the two-tailed test. Assuming $\alpha = .05$ and a critical value of -1.96, the null hypothesis cannot be rejected based on (Z = -1.391). In other words, no evidence exists that the sequence is non-random. If the

SPSS Output

Runs Test

	Major/ Non Major
Test Value	1.5000
Cases < Test Value	7
Cases >= Test Value	7
Total Cases	14
Number of Runs	5
Z	**-1.391**
Asymptotic. Sig. (2-tailed)	.164
Exact Sig. (2-tailed)	**.155**

a Median

SPSS Procedure:
Use data set: **Runs test**

1. Nonparametric
2. Runs
3. Move *Major/Non Major* from the "Variable List" window to the "Test variable list" window
4. In the "Cut point" window, check "median"
5. Click "OK"

question of interest were asked such that a directional or one-tailed test were appropriate, then the statistical significance would be (p = .0775) for this example. This type of hypothesis might be tested if the question of interest concerned whether there were a very small number of runs with all the majors (M's) completing the exam sooner than the non-majors (N's).

One Sample (Ordinary) Sign Test: Ordinal

One of the oldest nonparametric procedures, the Sign Test, has been traced from the 1700s. The data are converted to a series of plus and minus signs by subtracting the median or measure of interest from each observation. The test evaluates the number of plus signs and minus signs.

Assumptions:

1. The variable of interest is measured on at least an ordinal scale.
2. The variable of interest is continuous. The n sample measurements are designated by $X_1, X_2, \ldots, X_n$.
3. The sample is a random sample of independent measurements from a population with a median M that is hypothesized to be the median M_o.

Hypotheses:

Two-tailed: H_o: $M = M_o$, H_1: $M \neq M_o$,
One-tailed: H_o: $M \geq M_o$,H_1: $M < M_o$, or H_o: $M \leq M_o$, H_1: $M > M_o$

Procedure:

1. State the null and alternative hypotheses.
2. Select a level of significance α.
3. Record the sign of the difference obtained by subtracting the hypothesized median "M_o" from each sample value. (If the median of the sampled population (M) is actually M_o then there will be about the same number of plus signs as minus signs.)
4. If there is a sufficiently small number of plus or minus signs, reject the null hypothesis.

This test is based on the binomial distribution where the probability for a given type of sign is equal (p = .5). For the two-tailed test, the researcher would reject H_o at the a level if the probability of observing as few or fewer of the less frequently occurring sign in a random sample of size n is less than or equal to $\alpha/2$, i.e., (.025). For the one-tailed test, reject the H_o if the probability of observing as few or fewer of the appropriate sign in a random sample of size n is less than or equal to a, i.e., (.05). For the H_o of $M \leq M_o$, then with (X_i-M_o), H_o will be rejected if there are too many plus signs. For the H_o of $M \geq M_o$, then reject the H_o if there are too many minus signs.

Example: The value of the benefits package at an institution is 46% of the salary package. Researchers want to answer the following question: *Is the value of the benefits package at the institution (as a % of salary) consistent with that of similar institutions?*

Let $M_o = 46$ (the institution's benefit %), H_o: 46%

Obs.(n_i)	1	2	3	4	5	6	7	8	9	10	11
% (X_i)	25	22	45	34	39	23	15	36	44	46	50
X_i-46	-	-	-	-	-	-	-	-	-	0	+

SPSS Procedure:
Use data set: **One Sample Sign Test**

1. Nonparametric Tests
2. Binomial Test
3. Move *Score* to "Test Variable List"
4. Under "Define Dichotomy" - Choose "Cut point" and enter 46
5. "Test Proportion" equals .50
6. Click "OK"

Hypotheses (Two-tailed):
H_o: $M = M_o$, H_1: $M \neq M_o$, $\alpha = .05$

SPSS Output

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (2-tailed)
score	Group 1	<= 46	10	.91	.50	.012
	Group 2	> 46	1	.09		
	Total		11	1.00		

Conclusion: Reject the null hypothesis (H_o: $M = M_o$) because $p = .012 < .05$.

Test results reveal that the benefits package is not consistent with similar institutions' benefits packages. Note: if one thought initially the institution was higher (e.g. H_1: $M < M_o$ and H_o: $M \geq M_o$), this would have given a rejection region if one had too many minus signs for (M-M_o). The directional hypothesis would have given one-tailed rejection regions. The researcher would have rejected the null at the significance level of .006 and would have been very confident that the institution had a larger benefit package than the similar institutions. The tie for observation (n_{10}) is a problem and most likely occurred because of rounding. The SPSS handles this with "<=" but another approach is to leave the observation out and to do the computation with ten observations rather than eleven. For $N > 12$, the normal approximation works well with an expected value of .5*n and a standard deviation of $.5 * \sqrt{n}$.

Wilcoxon Signed Rank Test: Interval

The Wilcoxon Signed Rank Test is used to determine if a specific point or hypothesized median could be the population median. It is "applied as a one-sample median test by ranking the data and ... assigning a minus sign or plus sign to each rank assigned to the datum" (Zar, 1984, p. 114). In addition to the sign of the differences, this procedure uses magnitude of the difference of each observation from a hypothesized median.

Assumptions:

1. The variable is continuous.

2. The underlying or population distribution is symmetrical.
3. The scale of measurement is interval so that the observations can be placed in rank order.
4. The sample is a random sample of independent measures from a population with an unknown Median M that is compared with a hypothesized median M_o.

Hypotheses:

Two-tailed: H_o: $M = M_o$, H_1: $M \neq M_o$.
One-tailed: H_o: $M \geq M_o$, H1: $M < M_o$, or H_o: $M \leq M_o$, H_1: $M > M_o$

Procedure:

1. State the null and alternative hypotheses.
2. Select the level of significance α
3. Subtract the hypothesized median from each observation. For each observation find $D_i = X_i - M_o$. Eliminate any observation where the difference is zero.
4. Rank the differences from smallest to largest without regard to the signs (absolute value). If two differences are the same size, assign them the average of the two ranks
5. Assign each rank the sign (+ or -) of the difference (D_i).
6. Obtain the sum of the ranks with positive signs; call it T+. Obtain the sum of the ranks with negative signs; call it T-.
7. Compare the two "sum of ranks." If they are sufficiently different, then reject the hypothesis.

As noted above, unlike the previous Sign Test which uses the direction of the difference and not the magnitude of the difference, the Wilcoxon Signed Rank Test uses the magnitude in an ordinal measure as well as the difference. This increase in information causes the Wilcoxon Signed Rank Test to have much more statistical power than the Sign Test. It is assumed that the underlying population from which the sample is drawn is symmetrical about the hypothesized median M_o, leading to the assumption that the difference $(X_i - M_o)$ is symmetrical about zero. In this case, the median will equal the mean of the data.

Example: For the observations shown, researchers want to answer the following question: *Does the median IQ of the population from which these observations were drawn equal 107?*

Hypotheses (Two-tailed): H_o: $M = M_o$, H_1: $M \neq M_o$. $\alpha = .05$

Conclusion: Fail to reject the null hypothesis (H_o: $M = M_o$) because $p = .451 > .05$.

Observation	IQ	D = X - M	Rank	Signed Rank
1	99	-8	7	-7
2	100	-7	6	-6
3	90	-17	11	-11
4	94	-13	10	-10
5	135	+28	14	+14
6	108	+1	1	+1
7	107	0		
8	111	+4	5	+5
9	119	+12	9	+9
10	104	-3	4	-4
11	127	+20	13	+13
12	109	+2	2.5	+2.5
13	117	+10	8	+8
14	105	-2	2.5	-2.5
15	125	+18	12	+12
				T+ = 64.5 T- = 40.5

SPSS Procedure:
Use data set: **Wilcoxon Signed Rank Test**

1. "Nonparametric Tests"
2. "2 Related Samples Tests"
3. From "Variable List" select *Ho* and *iq1* and move to "Test Pair(s) List"
4. Under "Current Selections", "Variable 1" equals *Ho and* "Variable 2" equals *iq1*
5. Under "Test Type" choose "Wilcoxon"
6. Click "OK"

Based on these results we are unable to reject the null hypothesis that the median IQ of the population is 107 for ($\alpha = .05$). Note that SPSS drops the observation if there is a tie with the hypothesized median of 107. Also the subtraction occurs based on the sequence of variables in the data set where the first variable is subtracted from the second.

SPSS Output

Ranks

		N	Mean Rank	Sum of Ranks
IQ - Ho Median	Negative Ranks	6(a)	6.75	40.50
	Positive Ranks	8(b)	8.06	64.50
	Ties	1(c)		
	Total	15		

a. IQ < Ho Median
b. IQ > Ho Median
c. IQ = Ho Median

Test Statistics (b)

	IQ - Ho Median
Z	-.754(a)
Asymp. Sig. (2-tailed)	.451

a. Based on negative ranks.
b. Wilcoxon Signed Ranks Test

Tests of Location: Two Independent Samples

Median Test: Ordinal

The Median Test is used when evaluating whether two independent samples with the same distribution have the same median. It is similar to the signed test in that it uses only the direction of differences of observations for the two samples. This statistic considers the variation in both samples and determines the likelihood that the two samples come from the same distribution with the same median. This test considers a specific number (σ) and determines the proportion of the two samples that are above and below the specific

number. Though it does not have to be so, the number (σ) is generally the median of the two combined samples.

Assumptions:

1. The data consists of two independent random samples.
2. The first sample is from a population with median M_x and the second sample is from a population with the median M_y.
3. The measurement scale employed is at least ordinal.
4. The variable of interest is continuous.
5. The two populations from which the samples were drawn have the same shape.
6. If the two populations have the same median, then the probability p that an observation will exceed the specific number (σ) is the same for each sample.

Hypotheses:

Two-tailed: $H_o: M_x = M_y$, $H_1: M_x \neq M_y$.
One-tailed: $H_o: M_x \geq M_y$, H1: $M_x < M_y$, or $H_o: M_x \leq M_y$, $H_1: M_x > My$

Procedure:

1. Select and state the hypotheses.
2. Select the level of significance α.
3. Select a specific number σ. In general most select the grand median of the two combined samples because this is more likely to give about half above and half below for each group.
4. Compute the proportion of observations in each sample that are above and below the specific number σ.
5. Compare these proportions, and if they are sufficiently different in the two samples, conclude that the samples do not have the same median.

Given we compute a grand median; we can use this for the σ. We would then expect about half the observations from each sample to fall above the median and about half to fall below. Also note that results of counts are placed into a 2 x 2 contingency table as shown below. If a table has frequencies that are too small for a Chi-Square analysis, then the Fischer's exact test can be used (Zar, 1984, p. 71).

Definitions:

Relationship to Median	X	Y	Total
Above	a	b	a+b
Below	c	d	c+d
Total	$a+c = n_1$	$b+d = n_2$	$n_1 + n_2$

Example: The Department Head for English is concerned about recruiting new faculty and has asked Institutional Research to determine if a problem exists. There are thirty-two faculty members in his department and there are sixteen faculty members in a comparison department at another university. Researchers want to answer the following question: *Is the median faculty salary for the Department of English at this university equal to the median faculty salary at the closest comparison institution's department?*

Relationship to Median Salary	Your Institution (English Department)	Comparison Institution (English Department)	Total
Above	12	12	24
Below	20	4	24
Total	32	16	48

Hypotheses (Two-tailed): H_o: $M_x = M_y$, H_1: $M_x \neq M_y$. $\alpha = .05$

SPSS Output

Relationship to Median

Count

		Institution		Total
		Your Institution	Comparison Institution	
Relationship to Median	Above	12	12	24
	Below	20	4	24
Total		32	16	48

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	6.000(b)	1	.014	.030	.015
Continuity Correction(a)	4.594	1	.032		
Likelihood Ratio	6.207	1	.013	.030	.015
Fisher's Exact Test				**.030**	**.015**
Linear-by-Linear Association	5.875	1	.015	.030	.015
N of Valid Cases	48				

Conclusion: Reject the null hypothesis H_o: $M_x = M_y$ because $p = .03 < .05$. Based on the results of the Fisher's Exact Test, one would conclude

SPSS Procedure:
Use data set: **Median test confirmation**
Note: These are summarized data and as such you have to weight the data

1. Select "Data" on the main menu
2. From "Weight Cases" select "Weight Cases by"
3. Under "Frequency Variable" select *Salaries* from the "Variable List"
4. Click "OK"
5. In the lower right hand corner of the screen you should see "Weight On."

Analysis Procedure:
1. Under "Analysis" select "Descriptive Statistics"
2. Crosstabs"
3. From the "Variable List" move" *Relationship to Median* to the "Rows" category
4. From the "Variable List" move *Institution* to the "Columns" category
5. Click "Statistics" Tab at bottom of window
6. Select "Chi Square"
7. Click "Continue"
8. Under "Cell Display, Counts" choose "Observed"
9. Click "Continue"
10. Click "OK"

that the salaries in the department are not comparable to those at the other institution. In this case, there are too many salaries at the first institution below the grand median of the combined distribution of salaries.

Note that this test can also be obtained from (Nonparametric>K Independent Samples) where one can check the Median option, and identify the grouping variable as *Institution*, and identify that it has the value of "1" or "2".

Mann-Whitney U: Ordinal

The ability to consider the magnitude of the data as well as their direction once again provides a more powerful statistical test. When observations are given a rank within the data set that combines both samples, the average rank of the observations in one sample can be compared with the average rank of the observations in the other sample. If both samples have the same median, they should have approximately the same average rank.

The Mann-Whitney Test is a nonparametric test that is analogues to the two-sample parametric t-test (Zar, 1984). Its assumptions are based on the likelihood of various dichotomous patterns occurring when observations are ordered. If one can assume that the data are measured on any ordinal scale then the analyst can use rank statistics as a basis for the assumptions. The distribution of these ranks can then be used to compute the Mann-Whitney U. As will be seen later, the use of ranked statistics generalizes to the situation of where there are more than two samples. This is similar to how the Analysis of Variance generalizes to multiple samples while the t-test is restricted to two samples.

Assumptions:

1. The data consist of a random sample of observations from population 1 with unknown median M_x and another random sample of observations from population 2 with unknown median M_y.
2. The two samples are independent.

3. The variable observed is a continuous random variable.
4. The measurement scale employed is at least ordinal.
5. The distribution functions of the two populations do not differ. This means they have homogeneous variance.

Hypotheses:

Two-tailed: $H_o: M_x = M_y$, $H_1: M_x \neq M_y$

One-tailed: $H_o: M_x \geq M_y$, $H_1: M_x < M_y$, or $H_o: M_x \leq M_y$, $H_1: M_x > M_y$,

Procedure:

1. State the null and alternative hypotheses.
2. Select the level of significance α.
3. Assign a rank to all of the observations regardless of their group membership.
4. Compute the average rank for each group.
5. Compare the average ranks and evaluate the likelihood that a difference of the resulting magnitude and direction could occur by chance.

Example: Twenty-seven faculty members at an institution have been given a survey to determine their satisfaction with current course technology. Seventeen of the faculty members are in the Mathematics department and ten are in the History department. A high score represents a high level of satisfaction, with the highest level of satisfaction given a rank of one. Researchers want to answer the following question: *Is there a difference in the satisfaction scores of the faculty in the Mathematics department and the faculty in the History department?*

Hypotheses (Two-tailed): $H_o: M_x = M_y$, $H_1: M_x \neq M_y$ $\alpha = .05$

SPSS Procedure:
Use data set: **Mann-Whitney test**

1. "Nonparametric Tests"
2. "2-Independent Samples Tests"
3. Under "Test Variable List" select *Satisfaction Scores* from the "Variable List"
4. Under "Grouping Variable" select *Department* from "Variable List"
5. Under "Defined Groups": "Group 1" equals *1* and "Group 2" equals *2*
6. Click "Continue"
7. Under "Test Type" select "Mann-Whitney U"
8. Click "Ok"

SPSS Output

Ranks

	Department	N	Mean Rank	Sum of Ranks
Satisfaction Score	Mathematics	17	17.44	296.50
	History	10	8.15	81.50
	Total	27		

Test Statistics (b)

	Satisfaction Score
Mann-Whitney U	26.500
Wilcoxon W	81.500
Z	-2.938
Asymp. Sig. (2-tailed)	.003
Exact Sig. [2*(1-tailed Sig.)]	**.002(a)**

a Not corrected for ties.
b Grouping Variable: Department

Conclusion: Reject the null hypothesis (H_o: $M_x = M_y$) because $p = .002 < .05$.

Results of the tests indicate that there is a difference in the satisfaction levels of the two faculties ($p = .002$). Please note that this is a nondirectional test (2-tailed) and does not answer questions such as whether the History faculty are more satisfied than the Mathematics faculty or vice versa. If using a directional hypothesis, the null hypothesis would be rejected given the ($p = .001$) level using a one-tailed level of significance ($p = .05$). Note: The same results can be obtained if the "Test Variable" were *Rank* instead of *Satisfaction Score*. It is also interesting that tied ranks within a group do not make a difference in the statistic. However, a tie across groups will make a difference and is typically handled with the average rank procedure.

Tests of Location: Two Related Samples

McNemar Test: Nominal

One basic question that researchers might ask is whether or not a group has changed over time or as a result of some experience. For the case where there is a proportion for a group before and after some activity and the outcomes are dichotomous, a test for dependent portions should be used. The proportions being compared are based on just one group, or two related groups, rather than two independent groups. This is equivalent to asking whether the number that changed from one category to the other, for example from passing to failing, is the same as the number that changed in the other direction, from failing to passing. This reduces to a binomial test where the data can be summarized in a 2 x 2 contingency table. The changes in one direction are compared to total changes, where under the null hypothesis, an equal number of changes will be noticed in both directions.

Assumptions:

1. The data consists of n randomly selected subjects measured at two points in time or randomly selected pairs of subjects where the subjects are paired on some characteristic or set of characteristics.
2. The measurement scale is nominal with four categories that can be represented as yes-yes, yes-no, no-no, and no-yes for the two measures of each pair.
3. Let p_1 be the proportion of individuals with a "yes" characteristic on the first measure and p_2 be the proportion of individuals with a "yes" characteristic on the second measure.
4. The treatment does not change the proportion.

Hypotheses:

Two-tailed: H_o: $p_1 = p_2$, H_1: $p_1 \neq p_2$

One-tailed: H_o: $p_1 \leq p_2$, H_1: $p_1 > p_2$, or H_o: $p_1 \geq p_2$, H_1: $p_1 < p_2$

Procedure:

1. State the null and alternative hypotheses.
2. Select a level of statistical significance α.
3. Arrange the group of observations in the 2 x 2 contingency table.
4. Compare the proportions to see if they are sufficiently different to reject the null hypothesis.

Example: Students are administered two parallel forms of a test. The first test is given prior to participation in a math orientation program designed to improve success in basic math courses; the second following completion of the orientation program. Researchers want to know the answer to the following question: *Is participation in the orientation program related to the proportion of students who pass the test?*

SPSS Output

No Orientation * Orientation
Count

		After Orientation		
		Passed	Failed	Total
Before Orientation	Passed	26	15	41
	Failed	7	37	44
Total		33	52	85

Chi-Square Tests

	Value	Exact Sig. (2-sided)
McNemar Test		.134(a)
N of Valid Cases	85	
a. Binomial distribution used.		

SPSS Procedure:
Use data set: **McNemar Test for Two Related Samples**
Note: These are summarized data and as such you have to weight the data.

1. Select "Data" from the main menu
2. From "Weight Cases" select "Weight Cases by"
3. Under "Frequency Variable" select *Number of Students* from the "Variable List"
4. Click "OK"
5. In the lower right hand corner of the screen you should see "Weight On."

Analysis Procedure:
1. "Descriptive Statistics"
2. "Crosstabs"
3. From "Variable List" move *No Orientation* to the "Rows" category
4. From "Variable List" move *Orientation* to the "Columns" category
5. Click "Statistics" Tab at bottom of window
6. Select "McNemar"
7. Click "Continue"
8. Click "OK"

Hypotheses (Two-tailed):
H_o: $p_1 = p_2$, H_1: $p_1 \neq p_2$, $\alpha = .05$

Conclusion: Fail to reject the null hypothesis (H_o: $p_1 = p_2$) because $p = .134 > .05$.

Because the significance level is greater than .05, researchers cannot conclude that participating in orientation is related to passing Math 101. In other words, there appears to be no significant difference in the performance on the posttest following participation in the orientation session.

Sign Test for Matched Pairs: Ordinal

Sometimes a researcher has a set of data that represents pairs of measures across a set of individuals or activities. One of the basic questions

about such pairs of observations is: What is the relationship of one set of scores to the other set of scores? The *Sign Test for Matched Pairs* is a useful technique for answering this question. It enables the researcher to examine the differences between the pairs of scores and determines if one set of scores is larger than the other. It is an extension of the *One Sample Sign Test* described in the previous section. It is similar to a parametric paired "t-test" except that, as a nonparametric technique, it does not require the assumptions of normal distribution or linear scale.

Assumptions:

1. The data consist of a random sample of n pairs of measures where the pairing is based on some set of characteristics. The variable of interest is $X_i - Y_i = D_i$ where D_i has a + or - sign. The parameter about which we make inferences is M_d. This is the median of the differences between X and Y for the n pairs.
2. The n pairs of observations are independent from the other pairs of observations.
3. The measurement scale is at least ordinal within each pair so the largest can be determined.
4. The variable under study (D_i) is continuous which means there are no ties between the pairs of measures.

Hypotheses:

Two-tailed: $H_o: M_d = 0$, $H_1: M_d \neq 0$.

One-tailed: $H_o: M_d \leq 0$, $H_1: M_d > 0$, or $H_o: M_d \geq 0$, $H_1: M_d < 0$

Procedure:

1. State the null and alternative hypotheses.
2. Select the level of significance α.
3. For each pair, record the sign of the difference obtained by subtracting Y_i from X_i or by otherwise determining dominance.
4. Eliminate observations tied across the two measures.
5. If the two sets of scores come from the same population, there should be about as many plus signs as minus signs.

Example: There are ten departments in the College of Letters and Sciences. The Dean is interested in determining whether the college has a balanced instructional activity between Fall and Spring semesters. Instructional activity is measured as SCH/Faculty FTE. Researchers want to answer the following question: *Based on SCH/Faculty FTE ratios for Fall and Spring semesters, do the departments have a different productivity per FTE Faculty member in the Spring rather than the Fall?*

The following information is available:

Term	Dpt1	Dpt2	Dpt3	Dpt4	Dpt5	Dpt6	Dpt7	Dpt8	Dpt9	Dpt10
Fall	463	462	462	456	450	426	418	415	409	402
Spring	523	494	461	535	476	454	448	408	470	437
Sign (M_d)	+	+	-	+	+	+	+	-	+	+

Frequencies **SPSS Output**

		N
spring - fall	Negative Differences(a)	2
	Positive Differences(b)	8
	Ties(c)	0
	Total	10

a spring < fall
b spring > fall
c spring = fall

SPSS Procedure:
Use data set: **Sign Test _two related samples**

1. "Correlate"
2. "Bivariate Correlations"
3. From "Variable List" move *fall* and *spring* to "Variables" window
4. Under "Correlation Coefficients" select "Spearman"
5. Under "Test of Significance" select "Two-tailed"
6. Click "Flag significant correlations"
7. Click "OK"

Test Statistics (b)

	spring - fall
Exact Sig. (2-tailed)	.109(a)

a. Binomial distribution used.
b. Sign Test

Hypotheses (Two-Sided): H_o: $M_d = 0$, H_1: $M_d \neq 0$, $\alpha = .05$

Conclusion: Fail to Reject the null hypothesis (H_o: $M_d = 0$) because $p = .109 > .05$.

The hypothesis was stated as a two-tailed test (H_o: Spring = Fall). There is not sufficient evidence to reject the statement that instructional activity is the same in the Spring as in the Fall.

Wilcoxon Matched Pairs Signed-Rank Test

The Wilcoxon Matched Pairs Signed-Rank Test procedure is designed to determine if medians of two related measures are equal. Where the preceding Binomial Sign Test for Matched Pairs only required an ordinal scale within each pair, sometimes the analyst is able to assume that the data are measured on an interval scale. If the data can meet this additional assumption, a test that considers not only the direction of a difference but also the magnitude of the difference between pairs can be used. This increase in statistical power means that it will be much more likely to reject a null hypothesis when, in fact, it is false. For example, if the data from the preceding example for the Sign Test is used in this test, the researcher would be able to reject the Null Hypothesis ($p = .006$). In addition to the assumption that the data are measured on an interval scale, this test requires that the population of differences is symmetrical about their median.

In this test, the differences are first ranked based on their absolute

value. Signs are then assigned to the ranks based on the sign of the difference. The sum of the positive ranks is compared to the sum of the negative ranks. If the two distributions have the same median then the sum of the positive ranks should be about the same as the sum of the negative ranks.

This test is similar to the one sample Wilcoxon Signed Ranked Test just as the parametric paired t-test is similar to the one sample t-test. It is also helpful to remember that when the paired-sample t-test is applicable, the Wilcoxon paired sample test is applicable (Zar, 1984, p. 153).

Assumptions:

1. The data for the analysis consists of n values of the difference $D_i=Y_i-X_i$. Each pair of measurements is taken on the same subject or on subjects that have been paired on one or more attributes.
2. The sample of pairs is random with respect to the overall population of pairs.
3. The differences represent observations on a continuous random variable.
4. The distribution of the population of differences is symmetric about their median M_d.
5. The differences are independent.
6. The differences are measured on an interval scale.

Hypotheses:

Two-tailed: H_o: $M_d = 0$, H_1: $M_d \neq 0$

One-tailed: H_o: $M_d \leq 0$, H_1: $M_d > 0$, or H_o: $M_d \geq 0$, H_1: $M_d < 0$

Procedure:

1. State the null and alternative hypotheses.
2. Select the level of significance α.
3. For each observation find $D_i = Y_i - X_i$.
4. Eliminate any observation where the difference is zero.
5. Rank the differences from smallest to largest without regard to the signs (absolute value). If two differences are the same size, assign them the average of the two ranks.
6. Assign each rank the sign (+ or -) of the difference (D_i).
7. Obtain the sum of the ranks with positive signs; call it T+. Obtain the sum of the ranks with negative signs; call it T-.
8. Compare the two sums of ranks. If they are sufficiently different, then reject the hypothesis.

Example: A faculty member was concerned about the effect of stress on the learning situation for a class of nine students. She decides to enroll the students in a personal time management class. She administers parallel forms of a Stress Test before and after the time management class and

obtains the following scores. A high score represents a greater degree of stress. Researchers want to answer the following question: *Does a personal time management class significantly alter the score of students on a Stress Test?* (High scores reflect high stress, fifty items, and interval stress measure.)

The following information is available:

Student	Before Class (X_i)	After Class (Y_i)	$D_i = Y_i - X_i$	Signed Rank D_i
1	33	21	-12	-7
2	17	17	0	Omit
3	30	22	-8	-4
4	25	13	-12	-7
5	36	33	-3	-1
6	25	20	-5	-2
7	31	19	-12	-7
8	20	13	-7	-3
9	18	9	-9	-5

SPSS Output

Ranks

		N	Mean Rank	Sum of Ranks
after - before	Negative Ranks	8(a)	4.50	36.00
	Positive Ranks	0(b)	.00	.00
	Ties	1(c)		
	Total	9		

a. after < before
b. after > before
c. after = before

Test Statistics (b)

	after - before
Z	-2.533(a)
Asymp. Sig. (2-tailed)	.011

a. Based on positive ranks.
b. Wilcoxon Signed Ranks Test.

SPSS Procedure:
Use data set: **Wilcoxon Matched Pairs Test**

1. "Nonparametric Tests"
2. "2 Related Samples Tests"
3. From "Variable List" select *before* and *after* and move to "Test Pair(s) List"
4. Under "Current Selections", "Variable 1" equals *before* and "Variable 2" equals *after*
5. Under "Test Type" select "Wilcoxon"
6. Click "OK"

Hypotheses (Two-tailed):
$H_o: M_d = 0$, $H_1: M_d \neq 0$ $\alpha = .05$

Conclusion: Reject the null hypothesis ($H_o: M_d = 0$) because $p = .011 < .05$.

Results of a two-tailed test suggest that participation in the time management class alters student stress as measured by this instrument ($p = .011$). If using a one-tailed hypothesis, the order of the variables in the columns will impact the choice of directional tests. Interpretation thus becomes more difficult. Nevertheless, the negative Z score (based on subtraction of the "Before" score from the "After" score) suggests that stress was greater before participation in the class. This one-directional hypothesis can be

tested (H_o: $M_d \geq 0$, H_1: $M_d < 0$), and the decision will be made to reject the null hypothesis ($p = .005$).

Tests of Location: Three or More Independent Samples

Median Test: Ordinal

When there are more than two groups of independent measures, there are several tests to determine if the multiple groups could have come from the same underlying distribution. One of these methods is an extension of the *Median Test.* If the groups came from the same underlying population, they should all have about the same portion of numbers above some selected number (δ). This number is often selected to be the grand median of the combined groups because this tends to give a balance of observations above and below the number. The Median Test also works if responses are above or below some opinion or position such as Agrees/Disagrees or Likes/Dislikes. The Median Test uses the direction of the differences. Its exact likelihood is computed from a multivariate extension of the hypergeometric distribution. If there are at least twenty-five observations and at least five for each group, it can be approximated as a χ^2 Test of Independence with a 2 x k matrix. This procedure is discussed later in this chapter as the Chi Square Test of Independence for Two Independent Samples. More information can be found in Gibbons (1971, p. 196-198) or Sheskin (1997, p. 232-233).

Kruskal-Wallis One-way Analysis of Variance by Ranks

If the observations reflect an ordinal scale, they add information on both the magnitude of ranks as well as direction. This additional information can permit use of a much more powerful test called the Kruskal-Wallis One-way Analysis of Variance by Ranks. This test extends the Mann-Whitney U from two groups to more than two groups, much as the ANOVA extends the t-test. It does require a continuous scale where the *Median Test* only requires a dichotomy.

Assumptions:

1. The data for analyses consist of k random samples of sizes $n_1, n_2, n_3, \ldots n_k$.
2. The observations are independent both within and among samples.
3. The variable of interest is continuous.
4. The measurement scale is at least ordinal.
5. The populations are identical except for a possible difference in location for at least one population.

Hypotheses:

Two-tailed: H_o: $R_1 = R_2 = R_3 = \ldots R_k$

H_1: At least one of the equalities is violated.

Procedure:

1. State the null and alternative hypotheses.
2. Select the level of significance α.
3. Rank all of the observations in a combined ranking with 1 being the smallest number.
4. Compute the average rank for each group. If the groups come from the same population they should have about the same average rank.
5. If there is a significant overall difference, use a multiple comparisons procedure to look at where the difference occurs.

Example: A researcher is examining the applications for financial aid given to students in the various departments for the colleges of Arts & Sciences, Education, and Engineering. There are ten departments in Arts & Sciences and six each in Education and Engineering. All of the departments have about the same number of majors. Researchers want to answer the following question: *Are there differences in the median number of financial aid applications by college?*

The following information is available:

	Arts & Sciences		Education		Engineering	
Observation	Apps	Rank	Apps	Rank	Apps	Rank
1	262	4	465	16	343	10
2	307	7	501	18	772	20
3	211	3	455	15	207	2
4	323	8	355	11	1048	22
5	454	14	468	17	838	21
6	339	9	362	13	687	19
7	304	6				
8	154	1				
9	287	5				
10	356	12				
Sum of Ranks		69		90		94

SPSS Output

Ranks

	College	N	Mean Rank
Number of Applicants	Arts and Science	10	6.90
	Education	6	15.00
	Engineering	6	15.67
	Total	22	

Test statistics (a, b)

	Number of Applicants
Chi-Square	9.232
Df	2
Asymp. Sig.	**.010**

a. Kruskal Wallis Test
b. Grouping Variable: College

Hypotheses: H_o: $R_1 = R_2 = R_3$;
H_1: At least one of the equalities is violated.
a = .05

Conclusion: Reject the null hypothesis (H_o: $R_1 = R_2 = R_3$) because $p = .010 < .05$. At least one of the equalities is violated.

The significance level of less than .05 suggests that there is a difference between

SPSS Procedure:
Use data set: **Kruskal-Wallis One-way ANOVA by Ranks**

1. "Nonparametric Tests"
2. "K Independent Samples"
3. From "Variable List" move *Number of Applicants* to "Test Variable List"
4. From "Variable List" move *College* to "Grouping Variable"
5. Click "Define Range"
6. "Minimum" equals 1 and "Maximum" equals 3
7. Click "Continue"
8. Under "Test Type" click "Kruskal-Wallis H"
9. Click "OK"

the medians of at least two of the groups. To determine where the difference is, a multiple comparisons test is conducted.

Multiple Comparisons Test

The test of statistical significance of the difference between groups is based on the absolute difference in ranks. Given the absolute difference in average ranks of two groups, i and j, is the difference greater than $(z_b)^* \sqrt{N(N+1)(1/n_i+1/n_j)/12}$.

The colleges of Arts & Sciences and Education can be used to demonstrate.

The required absolute difference between the first and second group is: Difference = $(1.96)^* \sqrt{(22(22+1)^*(1/10+1/6)/12)}$ or 6.57.

The 6.57 can then be compared to the difference in the average ranks of these two groups, which is $|(69/10)-(90/6)| = 8.1$. Because 8.1 is greater than 6.57, it can be concluded that the difference between these two medians is significant (Sheskin, 1997, p. 402).

Tests of Location: Three or More Related Samples

Cochran Q: Nominal

When there are three or more related samples with the dependent variable evaluated on a dichotomous measure, Cochran's Q can be used to determine if there is a difference between two or more of the dependent measures. This situation might occur in cases where the dependent measures are three or more classes taken by a cohort of students. The question of interest becomes whether there is a significant difference between the courses in the proportion of students who pass each course. This is an extension of the McNemar Test to a situation where there are more than two dependent measures.

Assumptions:

1. The data consist of a set of n observations that are repeated on two or more dependent measures. The data can be thought of as a table where the rows are the sample and the columns are the Treatments.
2. The variable of interest is dichotomous.
3. The proportion of observations that are in a specific category is p_k for the k^{th} category.

Hypotheses: Ho: $p_1 = p_2 = p_3 = \ldots = p_k$

H_1: At least one of the equalities is violated

Procedure:

1. State the null and alternative hypotheses.
2. Select the level of significance α.

3. Score each subject for each measure as a dichotomy.
4. Compute the proportion in each group who were in the "successful" or dominant group. (If the groups come from the same population they should have about the same proportion of successes.)
5. If there is a significant overall difference, use a multiple comparisons procedure to examine where the difference occurs.

Example: The newly arrived Provost of an institution wants to determine if student success is linked to specific types of courses. She believes this can be partially evaluated by determining whether there is a difference in the pass rate for a cohort of students enrolled in math, English, and chemistry courses. Researchers want to answer the following question: *Is there a significant difference between the courses in the proportion of students who pass each course?*

The following information is available:

Student	Math	English	Chemistry
1	1	1	0
2	0	1	0
3	1	1	0
4	0	1	1
5	0	1	0
6	0	1	1
7	0	0	0
8	0	1	0
9	1	1	0
10	0	1	0
11	0	0	0
12	0	0	1
Where 0 = Failed, 1 = Passed			

Hypotheses: H_o: $p_1 = p_2 = p_3$; H_1: At least one of the equalities is violated

$\alpha = .05$

Conclusion: Reject the null hypotheses (H_o: $p_1 = p_2 = p_3$) because $p = .027 < .05$.

Results indicate that the null of equal proportions should be rejected ($p = .027$) and that there is a difference between at least two of the groups. Though the multiple comparisons procedure has not

SPSS Output

Frequencies

	Value	
	Failed	Passed
Math	9	3
English	3	9
Chemistry	9	3

Test Statistics

N	12
Cochran's Q	7.200(a)
df	2
Asymp. Sig.	**.027**

a. 1 is treated as a success.

SPSS Procedure:
Use data set: **Cochran's Q**

1. Nonparametric Tests
2. K-Related Samples
3. Move Math, English, and Chemistry from the "Variables List" window to the "Test Variables" window
4. Under "Test type" check Crochan's Q
5. Click "OK"

been performed in this case, visual evidence suggests that the students are much more likely to pass English than they are to pass the other two courses. As with the ANOVA alternative, there are ways to look at the pair-wise comparisons to include using pair-wise application of the McNemar Test. These are described in the previously mentioned Handbook by Sheskin (1997, p. 472-474).

Friedman Two-Way Analysis of Variance by Rank: Ordinal

While the sign test for matched pairs provides a way to test the comparability of paired observations for each subject, there are frequently three or more observations for each subject. In this situation, it is appropriate to use the Friedman Two-Way Analysis of Variance by Rank. In this test, the different objects are rank-ordered based on some ordinal characteristic within the set of objects. For example, programs can be ranked in terms of how students rate them on desirability. Institutions can also be the samples or units of measure, and the size of their academic departments can be ranked for a set of academic departments. The Friedman Two-Way Analysis of Variance by Rank is generally more powerful than the Cochran's Q because it uses the magnitude of differences as well as direction of differences. By contrast, Cochran's Q uses only values of a dichotomous variable to denote the direction of the differences.

Assumptions:

1. The data consist of *b* mutually independent samples of size *k*. The data can be thought of as a table where the rows are the individuals in the sample and the column is the item being ranked.
2. The variable of interest is continuous. This means there are no ties.
3. There is no interaction between the sample and the items being ranked.
4. Each subject can rank order the items based on some order of magnitude. The average rank of each item is M_k.

Hypotheses: H_o: $M_1 = M_2 = M_3 = \ldots = M_k$
H_1: At least one of the equalities is violated.

Procedure:

1. Select and state the hypotheses.
2. Select the level of significance α.
3. Rank each item in a row based on some ordinal characteristic.
4. Compute the average rank assigned to each object across the samples. (If each object came from the same population they should have about the same proportion of successes.)
5. If there is a significant overall difference, use a multiple comparisons procedure to determine where the difference occurs.

Example: The University President is interested in grant expenditures in specific departments and, in particular, whether the pattern of grant expenditures is equal to the expenditure patterns of the same departments at other institutions. He has asked Institutional Research to evaluate the situation. Researchers want to answer the following question: *Is there a difference in the median ranking of grant expenditures for the physics, geology, chemistry, and biology departments among your institution and seven peer institutions?*

The following information is available:

Institution	Physics	Geology	Chemistry	Biology
1	1	4	3	2
2	2	3	1	4
3	3	2	4	1
4	1	2	3	4
5	3	2	4	1
6	2	1	4	3
7	1	2	4	3
8	1	3	4	2
Total	14	19	27	20

Note: While the underlying measures (grant expenditures) may be ratio or interval, you must rank for each observation measures within the separate institutions.

Ranks **SPSS Output**

	Mean Rank
Physics	1.75
Geology	2.38
Chemistry	3.38
Biology	2.50

Test Statistics (a)

N	8
Chi-Square	6.450
df	3
Asymp. Sig.	**.092**

a. Friedman Test

Hypotheses (Two-tailed): H_o: $M_1 = M_2 = M_3 = \ldots = M_8$;
H_1: At least one of the equalities is violated.

Conclusion: Fail to reject the null hypothesis (H_o: $M_1 = M_2 = M_3 = \ldots = M_8$) because $p = .092 > .05$.

SPSS Procedure:
Use data set: **Friedman Two-Way ANOVA by Rank**

1. "Nonparametric Tests"
2. K-Related Samples
3. From "Variable List" move all variables to the "Test Variables" window
4. Under "Test Type" choose "Friedman"
5. Click "OK"

Results of the tests suggest that there is no significant difference in the rankings of departments on grant expenditures among the institutions of interest. As such, there is no reason to perform a multiple comparisons test to determine which institutions differ. Also keep in mind that the findings do not suggest that there are no differences with respect to amounts spent by departments among the institutions.

If the null hypotheses had been rejected (a significant difference was

found in the above test), one would have concluded that there are some inequalities for which the differences are statistically significant. The difference between ranks can be tested using: $|R_j - R_i| >= z_b * \sqrt{(b*k*(k+1)/6)}$. Use $p = \alpha / [k(k-1)]$ for the level of a and obtain z for the probability level p from the normal distribution table, where b = observations and k = treatments or measures (Sheskin, 1997, p. 458-462).

Goodness of Fit: One Sample

Chi Square Goodness of Fit Test: Nominal

Sometimes the purpose of a research project is to determine if the frequencies for a set of categories are reasonably similar to what one would expect by chance when the frequencies come from a known distribution. In this case, if the categories are nominal — that is they have no natural order — then the appropriate test is called the *Chi Square Goodness of Fit Test.* This test is based on the Chi Square distribution. If there are k categories with specified numbers in the various cells, the Chi Square distribution of interest is the one which has (k-1) degrees of freedom. The Chi Square test statistic is computed as: $\chi^2 = \Sigma[(O_i - E_i)^2 / E_i]$, where O is the observed frequency, E is the expected frequency, and i is the category.

Researchers using the Chi Square should be aware of several caveats. The first is that the expected values of the cells need to be moderately large. One rule of thumb is that no cell should have expected values of less than one and 20% of the cells should not have values less than five. Frequencies that are too small tend to inflate the Chi Square, increasing the likelihood of rejecting the null hypothesis. The traditional process for the case where frequencies are small is to combine cells in some experimentally meaningful manner. The second caveat is that in cases where frequencies are small and there are only two categories, it may be appropriate to make an adjustment to reduce the computed Chi Square before it is compared with the table value. This is known as Yates' correction for continuity.

Assumptions:

1. The data available for analysis consist of a random sample of *n* independent observations.
2. The measurement scale may be nominal.
3. The observations can be classified into *r* non-overlapping categories that exhaust all classification possibilities. The categories are mutually exclusive. The number of observations falling into a given category is called the observed frequency of that category.

Hypotheses:

$H_o: O_i = E_i$; H_1 $O_i \neq E_i$

Procedure:

1. State the null and alternative hypotheses.
2. Select the level of significance α.
3. Obtain the frequencies of the categories for the observed distribution O.
4. Compute the expected frequencies based on $f(x_i)*N$ where $N = \Sigma O_i$, and $f(x_i)$ is the hypothetical distribution.
5. Calculate the $\chi^2 = \Sigma[(O_i-E_i)^2/E_i]$ and compare to the value needed to reject the null hypothesis.

Example: Students majoring in nutrition are required to select one of six minors in the College of Arts and Sciences. Researchers want to answer the following question: *Is the choice of minors for undergraduate students majoring in nutrition equally distributed among the six alternative programs in the College of Arts and Sciences?*

The following information is available:

Preferred Minor	Expected Frequency	Observed Frequency
Math	6	12
English	6	6
Biology	6	1
Chemistry	6	3
History	6	11
Psychology	6	3
Total	36	36

SPSS Procedure:
Use data set: **Goodness of Fit-One Sample**
These are summarized data and as such you have to weight the data.

1. Select "Data" from the main menu
2. From "Weight Cases" select "Weight cases by"
3. Under "Frequency Variable" select *Observed Frequency* from the "Variable List"
4. Click "OK"
5. In the lower right hand corner of the screen you should see "Weight On."

Analysis Procedure:
1. "Nonparametric Tests"
2. "Chi Square"
3. From "Variable List" move *Preferred Minor* to "Test Variable List"
4. Under "Expected Range" select "Get from data"
5. Under "Expected Values" select "Values"
6. Enter *6* for each Minor (You should see six 6s in this window.)
7. Click "OK"

SPSS Output

Preferred Minor

	Observed N	Expected N	Residual
Math	12	6.0	6.0
English	6	6.0	.0
Biology	1	6.0	-5.0
Chemistry	3	6.0	-3.0
History	11	6.0	5.0
Psychology	3	6.0	-3.0
Total	36		

Test Statistics

	Preferred Minor
Chi-Square(a)	17.333
df	5
Asymp. Sig.	.004

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 6.0.

Calculation:

$\chi^2=[(12-6)^2 + (6-6)^2 + (1-6)^2 + (3-6)^2 + (11-6)^2 + (3-6)^2]/6$

$\chi^2 = 17.333$ with 5 degrees of freedom

Hypotheses: H_o: $O_i = E_i$;
H_1: $O_i \neq E_i$

$\alpha = .05$

Conclusion: Reject the null hypothesis (H_o: $O_i = E_i$) because $p = .004 < .05$.

The results suggest that the choice of minors for nutrition majors is not equally distributed among the six programs from which students can choose. Please observe that there is caution generated by SPSS concerning the problem associated with the test if there are several cells with a low frequency. If the observed frequency is 0, SPSS will delete the category, which can cause an incorrect estimate.

Kolmogorov-Smirnov One Sample Test: Ordinal

The Chi Square Test uses only information about the number of observations in each category. It does not consider situations where the categories have a natural order. This means that the Chi Square Test is much less likely to identify a difference between the observed frequency and the expected frequency for ordinal categories. In order to take ordinal categories into account, it is necessary to use the *Kolmogorov-Smirnov One Sample Test.* This test computes the difference between the cumulative observed and expected frequencies and then compares the largest value of this cumulative difference to what might be expected by chance. This test is based on order statistics. While it requires a continuous distribution for its derivation it can be computed on functions that are discrete, in which case it becomes somewhat conservative. Its advantage over the *Chi Square Test* is that it can be computed on rather small frequencies where the *Chi Square Test* requires rather large sample sizes. In addition, it is an exact test where the *Chi Square Test* is actually a Chi Square distribution only as the sample becomes very large.

Assumptions:

1. The data are measured on a continuous and ordinal scale.
2. The data consist of the independent observations X_1, X_2 ...Xn, constituting a random sample of size *n* from some unknown distribution.
3. At any given point, the absolute difference in the frequencies at that point between what is observed and what is expected for a cumulative distribution f(x) is represented by T_i.

Hypotheses:

H_o: T_i is $= T_{max}$,

where T_i is the absolute difference between the expected cumulative frequency and the actual cumulative frequency at point "i" and T_{max} is the largest difference that would be expected by chance.

H_1: T_i is $> T_{max}$ for some "I".

Procedure:

1. State the null and alternative hypotheses.
2. Select the level of significance α.
3. Identify an expected distribution f(x).
4. Obtain the frequencies of the categories for the observed distribution and create a cumulative distribution where the sum for the category equals the sum of those frequencies plus frequencies in all preceding categories.
5. Compute the expected frequencies for the categories of the expected (hypothesized) distribution in a similar fashion.
6. Subtract the expected cumulative from the observed.
7. Determine if the absolute difference (T_i) is greater than would be expected by chance.

Example: Students entering your college are required to take a Math Placement Test. Four hundred and six students took the exam prior to the beginning of the Fall semester. The department head in mathematics is concerned about the effectiveness of the placement program for class scheduling purposes. He thinks the distribution of abilities should follow a normal distribution. Researchers want to answer the following question: *Are the actual math placement scores of entering first year students normally distributed?*

The department head has also recorded the scores after rounding to an integer and wants to know if that digit is normally distributed. Researchers must also answer the second question: *Are the math placement scores of entering first year students recoded as an integer normally distributed?*

Hypotheses (for both questions): H_o: T_i is $\leq T_{max}$; H_1: T_i is $> T_{max}$, $\alpha = .05$

SPSS Output

One-Sample Kolmogorov-Smirnov Test

		Math Placement Test	Math Placement Rounded
N		406	406
Normal Parameters(a,b)	Mean	15.495	15.63
	Std. Deviation	2.8210	2.821
Most Extreme Differences	Absolute	.047	.101
	Positive	.047	.101
	Negative	-.032	-.079
Kolmogorov-Smirnov Z		.951	2.034
Asymp. Sig. (2-tailed)		**.326**	**.001**

a Test distribution is Normal.
b Calculated from data.

SPSS Procedure:
Use data set: **Test for Normality**

1. "Nonparametric Tests"
2. "1-Sample K-S"
3. From "Variable List" move Math Placement Test and Math Placement Rounded to "Test Variable List"
4. Under "Test Distribution" select "Normal"
5. Click "OK"

Question 1: Conclusion: Fail to reject the null hypothesis (H_o: T_i is $= T_{max}$) because $p = .326 > .05$).

Question 2: Conclusion: Reject the null hypothesis (H_o: T_i is $= T_{max}$) because $p = .001 < .05$.

Failure to reject the null hypothesis associated with the first question suggests that it can be concluded that the Math Placement scores are normally distributed. By contrast, the null hypothesis associated with the second question was rejected. This suggests, based on the rounded score, that the scores do not come from a normal distribution ($p = .001$). It appears that the math test scores represent a normal distribution while the rounded scores do not.

Goodness of Fit: Two Independent Samples

Kolmogorov-Smirnov Two Sample Test: Ordinal

If one wants to compare two distributions with each other, the Kolmogorov-Smirnov statistic can be generalized from a one-sample test to a two-sample test. In the latter case, the statistical test is computed by comparing the empirical distribution functions of the two samples. As with the one-sample case, the two-sample case is based on order statistics, and it is necessary that the underlying distribution be continuous. Otherwise, the test becomes somewhat conservative. The distributions must also be measurements that reflect at least an ordinal scale.

Assumptions:

1. There are two distributions on a continuous and ordinal scale.
2. If the two distributions come from the same underlying population, then the difference between their cumulative distributions will not exceed some maximum amount. The cumulative distributions are compared at each level instead of comparing one of the distributions to a theoretical distribution.

Hypotheses:

H_o: $D_{m,n}$ is $\leq D_{max}$

where $D_{m,n} = \max|S_m(x) - T_n(x)|$ over values of x, $S_m(x)$ and $T_n(x)$ are the two cumulative distributions, and D_{max} is the largest difference that would be expected by chance.

H_1: $D_{m,n}$ is $> D_{max}$

Procedure:

1. State the null and alternative hypotheses.
2. Select the level of significance α.
3. Obtain the frequencies of the categories for the observed distributions and create cumulative distributions where the sum for the category equals the sum of those frequencies plus frequencies in all preceding categories...
4. Subtract one cumulative from the other and determine the largest difference.
5. Determine if the cumulative difference is greater than would be expected by chance.

Example: The Dean of the College of Arts and Sciences has observed varying levels of participation in service work among faculty in the departments of Mathematics and History. The Associate Dean suggested that this might in some way be related to job satisfaction. In response, a survey was administered to determine the job satisfaction of faculty in the two departments. Researchers want to answer the question: *Are the distributions of job satisfaction scores between the faculty of the Math and History departments equal?*

SPSS Output

Frequencies

	Department	N
Satisfaction Score	Mathematics	17
	History	10
	Total	27

Test Statistics (a)

		Satisfaction Score
Most Extreme Differences	Absolute	.706
	Positive	.000
	Negative	-.706
Kolmogorov-Smirnov Z		1.771
Asymp. Sig. (2-tailed)		.004

a. Grouping Variable: Department

SPSS Procedure:
Use data set: **Mann-Whitney test**

1. "Nonparametric Tests"
2. "2-Independent Samples Tests"
3. From "Variable List" move *Satisfaction Score* to "Test Variable List"
4. From "Variable List" move *Department* to "Grouping Variable"
5. Under "Defined Groups": "Group 1" equals *1* and "Group 2" equals *2*
6. Click "Continue"
7. Under "Test Type" select "Kolmogorov-Smirnov Z"
8. Click "Ok"

Hypotheses: H_o: $D_{m,n}$ is $\leq D_{max}$
H_1: $D_{m,n}$ is $> D_{max}$
$\alpha = .05$

Conclusion: Reject the null hypothesis (H_o: $D_{m,n}$ is $= D_{max}$) because $p = .004 < .05$.

In this case, it does not seem to be reasonable to claim that these two distributions come from a single underlying distribution. The null hypothesis is rejected (p = .004). It should be recalled that when these data were used to calculate the Mann-Whitney Test, the medians of the two distributions were found to be significantly different ($p < .003$).

Measures of Association: Two Variables

Measures of association concern both the amount and direction of association for multiple variables. They do not, in and of themselves, identify the likelihood that the association occurred by chance. It should be noted that association is not the same as agreement. For example, perfect association may well go with perfect disagreement. In this case, some of the measures will have negative signs.

2 x 2 Associations (Biserial, Point Biserial, Phi Coefficient, Techrachoric): Nominal

When the data are represented by one or two dichotomous variables, the question concerns whether the dichotomies are true dichotomies or whether the measures are continuous measures that were structured into dichotomies. There are four primary measures of statistical association that can be computed for two dichotomous variables. Definitions for these measures are presented below. The distributions of all of these variables are assumed symmetrical in that there is no causal order attributed to the measures. When the data are represented by one continuous variable and the other measure is a true dichotomous measure, this is computed as one would compute the Pearson Product Moment Correlation, a parametric equivalent.

Types of Statistical Association for Two Dichotomous Variables:

- *Biserial correlation coefficient* — r_{bi}
  - For use when one variable is continuous and the other is a dichotomous variable that reflects an underlying normal distribution.
- *Point biserial coefficient* — r_{pb}
  - For use when one variable is continuous and the other is a 'true' dichotomous variable.
- *Phi coefficient* — φ
  - For use with two 'true' dichotomous variables.
- *Tetrachoric correlation coefficient* — r_{tet}
  - For use with two artificial dichotomies where the variables have underlying normal distributions. (See Ender, 2004)

In this chapter, we demonstrate the Phi (φ) coefficient. We present this example as it uses the weakest set of assumptions. In other words, it makes no requirement about the underlying distribution of scores. It should be noted that in general it does not range from ± 1.0 but has a φ_{max}. It is computed using the Pearson Product Moment Correlation formula.

Hypotheses:

Two-tailed: H_o: $\rho = 0$, H_1: $\rho \neq 0$

One-tailed H_o: $\rho \leq 0$, H_1: $\rho > 0$, or H_o: $\rho \geq 0$, H_1: $\rho < 0$

Procedure:

1. State the null and alternative hypotheses.
2. Select the level of significance α.
3. Categorize the data into the four cells.
4. Compute the correlation.
5. Compare Φ to the value needed to reject the null hypothesis.

Example: A faculty member in the Math Department is interested in gender differences among students on successful outcomes, i.e., a passing score, for Math 206. Data are available to examine these differences given that those taking the Math 206 final exam are identified by gender. Researchers want to answer the question: *Is there a relationship between gender and passing the Math 206 final examination?*

SPSS Output

Gender * Outcome Crosstabulation

Count

		Outcome		
		Passed	Failed	Total
Gender	Male	23	10	33
	Female	5	2	7
Total		28	12	40

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	**Phi**	**-.014**	**.928**
	Cramer's V	.014	.928
N of Valid Cases		40	

a Not assuming the null hypothesis.
b Using the asymptotic standard error assuming the null hypothesis.

SPSS Procedure:
Use data set: **Phi Coefficient**
These are summarized data and as such you have to weight the data.

1. Select "Data" from the main menu
2. From "Weight Cases" select "Weight cases by"
3. Under "Frequency Variable" select *Number of Students* from the "Variable List"
4. Click "OK"
5. In the lower right-hand corner of the screen you should see "Weight On."

Analysis Procedure:
1. "Descriptive Statistics"
2. "Crosstabs"
3. From the "Variable List" move *Outcome* to "Rows" window
4. From the "Variable List" move *Gender* to "Columns" window
5. Click "Statistics Tab"
6. Under "Nominal" select "Phi and Cramer's V"
7. Click "Continue"
8. Click "OK"

Hypotheses (Two-tailed): H_o: $\rho = 0$, H_1: $\rho \neq 0$; $\alpha = .05$
Conclusion: Fail to reject the null hypothesis (H_o: $\rho = 0$) because $p > .928 > .05$.

Results of the test suggest that there is no significant relationship between successful outcomes for Math 206 and gender. Recall that the Phi Coefficient is for use with two dichotomous variables, in this case "passing" (1=passed; 0=not passed) and "gender" (1=male; 0=not male).

Chi Square Test of Independence - Two Independent Samples: Nominal

When there are two measures and both are categorical measures, one of the most common of all nonparametric statistics is the *Chi Square Test of Independence*. This test uses the Chi Square distribution to determine if two variables are significantly related. It should be noted that this is a measure of association and not agreement. Chi Square is computed by summing the squared differences in observed and expected frequencies after dividing by the expected frequency. The expected frequency is computed based on the marginal frequencies. If $f_{.j}$ is the sum of the frequencies in the j^{th} column and $f_{i.}$ is the sum of the frequencies in the i^{th} row and N is the total number of observations, then the expected frequency in the cell x_{ij} is $E_{ij} = (f_{.j} * f_{i.})/N$.

The computed χ^2 has (c-1)*(r-1) degrees of freedom, where "c" is the number of columns and "r" is the number of rows. The same cautions exist for this test as for the earlier discussion of using the Chi Square distribution

to test the similarity of a set of frequencies to a hypothesized distribution. Small expected cell frequencies inflate the computed statistic.

Assumptions:

1. The data consist of a simple random sample of size n from some population of interest.
2. The observations in the sample are cross-classified according to two criteria, so that each observation belongs to one and only one category of each criterion. The criteria are the variables of interest in a given situation.
3. The variables may be inherently categorical, or they may be quantitative variables whose measurements are capable of being classified into mutually exclusive numerical categories.

Hypothesis: $H_o: \chi^2 = 0$; $H_1: \chi^2 \neq 0$.

Procedure:

1. State the null and alternative hypotheses.
2. Select the level of significance α.
3. Categorize the data into the cells.
4. Compute the χ^2 and the degrees of freedom.
5. Compare χ^2 to the value needed to reject the null hypothesis with (c-1)*(r-1) degrees of freedom.

SPSS Output

Parental Monthly Income * College major Crosstabulation

			College major			
			Humanities and Social Sciences	Engineering	Agriculture	Total
Parental Monthly Income	< $3,000	Count	186	38	35	259
		Expected Count	153.7	53.1	52.2	259.0
	$3,000-$4,900	Count	227	54	45	326
		Expected Count	193.4	66.9	65.7	326.0
	$5,000-6,900	Count	219	78	78	375
		Expected Count	222.5	76.9	75.6	375.0
	$7,000-$9,900	Count	355	112	140	607
		Expected Count	360.2	124.5	122.3	607.0
	$10,000 +	Count	653	285	259	1197
		Expected Count	710.2	245.5	241.2	1197.0
Total		Count	1640	567	557	2764
		Expected Count	1640.0	567.0	557.0	2764.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	47.892(a)	8	.000
Likelihood Ratio	49.087	8	.000
Linear-by-Linear Association	32.681	1	.000
N of Valid Cases	2764		

a 0 cells (.0%) have expected count less than 5. The minimum expected count is 52.19.

Example: A researcher has collected a data set from the entering first year class that includes their intended major and also the income of their parents. She wants to answer the question: *Is there a relationship between parental income and the college major choice for entering first year student?*

Hypothesis: $H_o: \chi^2 = 0$; $H_1: \chi^2 \neq 0$; $\alpha = .05$

Conclusion: Reject the null hypothesis ($H_o: \chi^2 = 0$) because $p = .000 < .05$.

The researcher can conclude that a relationship exists between college major selections of entering first year students and parental monthly income. The residual is computed as (Observed minus Expected) so that a negative residual means that more would have been expected by chance with independent measures than were actually observed. The reverse is true for a positive residual.

SPSS Procedure:
Use data set: **Chi-Square Test of Independence**
These are summarized data and as such you have to weight the data.

1. Select "Data" from main menu
2. From "Weight Cases" select "Weight cases by"
3. Under "Frequency Variable" select *freq* from the "Variable List"
4. Click "OK"
5. In the lower right hand corner of the screen you should see "Weight On."

Analysis Procedure:
1. "Descriptive Statistics"
2. "Crosstabs"
3. From "Variable List" move *Parental Income* to "Rows" window
4. From "Variable List" move *College Major* to "Columns" window
5. Select "Statistics" tab
6. Check "Chi-square"
7. Click "Continue"
8. Select "Cells" tab
9. Under "Counts" select "Observed" and "Expected"
10. Under "Residuals" select "Unstandardized"
11. Click "Continue"
12. Click "OK"

Spearman – Rank Correlation Rho (ρ): Ordinal

When there are pairs of observations, one from set S and the other from set T, it is possible to compute the association between the two sets represented by the pairs. It is only necessary to rank the scores within each set. The rank correlation is based on the difference between the ranks and can range from plus one for perfect agreement to minus one for perfect disagreement. It has the expectation of zero when the two sets are independent. For ten pairs or more, the variance of the correlation is estimated by (1/(n-1)).

Assumptions:

1. The data consist of a random sample of n pairs of numeric or non-numeric observations.
2. Each pair of observations represents two measurements taken on the same object or individual.
3. These measures are on an ordinal scale. The data for a variable are ranked relative to all other observations for that variable from smallest to largest. The smallest observation is ranked 1.
4. If ties occur, the tied measures are typically assigned the mean of the tied ranks.

Hypotheses:

Two-tailed: H_o: $\rho = 0$, H_1: $\rho \neq 0$

One-tailed: H_o: $\rho \leq 0$, H_1: $\rho > 0$, or H_o: $\rho \geq 0$, H_1: $\rho < 0$

Procedure:

1. State the null and alternative hypotheses.
2. Select the level of significance α.
3. Rank the observations in each set.
4. Compute the differences in ranks and use this to compute the correlation.
5. Compare ρ to the value needed to reject the null hypothesis.

Example: There are ten departments in the College of Arts and Sciences. The researcher wants to answer the question: *Are the number* of credit hours generated by these departments in the Fall semester related to the number generated in the Spring semester?

SPSS Output

Correlations

			fall	spring
Spearman's rho	fall	Correlation Coefficient	1.000	.717(*)
		Sig. (2-tailed)	.	.020
		N	10	10
	spring	Correlation Coefficient	.717(*)	1.000
		Sig. (2-tailed)	.020	.
		N	10	10

* Correlation is significant at the 0.05 level (2-tailed).

SPSS Procedure:
Use data set: **Sign Test _two related samples**

1. "Correlate"
2. "Bivariate Correlations"
3. From "Variable List" move *fall* and *spring* to "Variables" window
4. Under "Correlation Coefficients" select "Spearman"
5. Under "Test of Significance" select "Two-tailed"
6. Click "Flag significant correlations"
7. Click "OK"

Hypotheses (Two-tailed): H_o: $\rho = 0$, H_1: $\rho \neq 0$, $\alpha = .05$

Conclusion: Reject the null hypotheses (H_o: $\rho = 0$) since $p = .02 < .05$.

Results of the test suggest that there is a relationship between the number of hours generated in the Fall and Spring semesters. The null hypothesis against which the test is run is for "no relationship". Note that the manner in which the question was asked —"are related" — suggests the use of a two-tailed test.

Measures of Association: Three or More Variables

Kendall's Coefficient of Concordance W: Ordinal

While *Spearman's Rank Order Correlation* works with two measures, it does not generalize well to more than two. A procedure similar to Spearman's computes the concordance of various measures. Concordance is defined as

having pairs of measures that are in the same direction. In other words, if ($x_i > x_j$) and the same relationship exists for the pair of observations on a second variable ($y_i > y_j$), then the pairs are concordant. If there are only two measures, the resulting statistic is frequently referred to as tau. When there are more than two measures, the statistic is referred to as *Kendall's W or Coefficient of Concordance.* This measure shows the amount of agreement for a set of rankings. The W statistic ranges from near 1.0 to 0.0 but never reaches these limits. It does not fall below zero since the concept of disconcordance does not have a definition with three or more rankings. As mentioned earlier, W is related to *Friedman's Two Way Analysis of Variance of Ranks* much as the Correlation Ratio Eta Squared (η^2) is related to ANOVA, where W explains the association and *Freedman's Two Way Analysis of Variance of Ranks* explains the differences in the average ranks.

Assumptions:

1. The data consist of b complete sets of observations or measures on k items.
2. The measurement scale is at least ordinal for the k items.
3. The observations as collected consist of ranks or are converted to ranks within the *k* observations.

Hypotheses:

H_o: $W = 0$;
H_1: $W \neq 0$.

Procedure:

1 State the null and alternative hypotheses.
2 Select the level of significance α.
3 Rank the "n" objects for each situation.
4 Compute the W.
5 Compare W to the value Q needed to reject the null hypothesis.

Example: Administrators in higher education are increasingly interested in comparing their institution with a group of peer institutions. The President of the researcher's institution wants to compare the relative positions (rankings) of a set of peer institutions (including his own) on three measures — number of applicants, faculty salaries, and research expenditures. Researchers want to answer the question: *Are the rankings produced by the three measures; Number of Applicants, Faculty Salaries, and Research Expenditures, significantly related to each other – and if so – how strongly are they related?*

Hypotheses: H_o:$W = 0$; H_1: $W \neq 0$; $\alpha = .05$

Note: To build the second data set, each case (e.g. row in the input data) is

SPSS Output

Case Summaries (a)

Peer Institutions	Rank of Applications	Rank of Average Salaries	Rank of Research Expenditures
1	6	8	6
2	1	7	7
3	3	4	2
4	5	3	4
5	8	6	8
6	2	1	1
7	4	2	3
8	7	5	5

a. Limited to first 100 cases.

Ranks

	Mean Rank
Institution 1	6.67
Institution 2	5.00
Institution 3	3.17
Institution 4	4.00
Institution 5	7.33
Institution 6	1.33
Institution 7	3.00
Institution 8	5.50

SPSS Procedure:
Use data set: **Kendall's Coefficient of Concordance Wb**

1. "Nonparametric Tests"
2. "K-Related Samples
3. Move *Inst1-Inst8* variables to "Test Variable" window
4. Check "Kendall's W"
5. Click "OK"

Test Statistics

N	3
Kendall's W(a)	.676
Chi-Square	14.195
df	7
Asymp. Sig.	.048

a Kendall's Coefficient of Concordance

the set of ranks and each variable is an item being judged. In this example, the first case is Applications, the second case is Average Salaries, and the third case is Research Expenditures. Each institution then becomes a variable (column in the input data).

Conclusion: Reject the null hypothesis (H_o: W = 0) because $p = .048 < .05$.

Results of the test suggest that the rankings for the eight institutions are statistically similar for the three measures. This finding indicates that the three measures are statistically related to each other and as such give comparable results. It should be noted that the test of significance is the same as the test that the eight schools came from the same population (Friedman's Two Way Analysis of Ranks).

Beyond Nonparametrics: Some Advanced Topics

While the preceding discussion focused some of the more traditional methods of non-parametric statistics, there are also other powerful tools that require fewer assumptions than the parametric tools. While the full demonstration and explanation of these tools are beyond the scope of this chapter, the following describes some of them so that those who see them as useful can follow up on their value through the included references.

OLAP

The emerging methodology of On Line Analytical Processing is basically a non-parametric methodology. It involves the data mining of events that are non-parametric. It is not assumed that there is an underlying distribution. Furthermore it is not assumed that there is a linear or higher level of the dependent measure. Within this methodology, there is an emerging set of analytical techniques (Thierauf, 1997).

- The first set of data mining tools model what is known as "neural networks." These methods are based on collections of inputs and outputs, and processing. Each node is capable of learning as it has a mechanism that allows it to learn pattern recognition. This pattern typically involves non-additive events that produce an expected outcome.

- "Decision trees" are the second type of data mining. They divide the data into groups based on values of variables. Typically the decision trees have algorithms that allow them to maximize the differences between groups compared to the variances within groups.

- "Rule induction" is the third type of data mining tools. These tools work to develop a set of if-then statements. These rules are typically many-to-one relationships although in their more complex forms they can approach many-to-many. As such they are similar to the learning process of neural networks.

- "Data visualization software" is the fourth and final type of data mining tools. These tools provide the ability to visualize up to as many as four variables in a single picture. Their effectiveness depends on the knowledge of the viewer about the capability of the tool and provider to show meaningful relationships.

It should be noted that these methods may or may not depend on a linear scale. Their non-parametric nature is on their use to produce results without developing an assumption of the underlying distribution of the observations.

Log-Linear Analysis

Log-linear analysis provides the opportunity to look at the independence of three or more variables based on the number/frequency of occurrences for the individual categories. This is an extension of the situation where the Chi Square Test of Independence would be useful except there are more than two independent sets of categories. The dependent variable, in the perspective of ANOVA, is the frequency or the count of occurrences that occur within the presence of the multiple independent variables. This methodology does assume certain underlying distributions, but in general is much less restrictive than the Analysis of Variance and other similar techniques. It does involve the use of the Chi Square frequency to estimate the failure of the model to fit the observed data. As such, it uses the Chi Square and some other metrics to measure a drop in the ability of a model to fit the data. If alternatives are nested, then the move to a more parsimonious model can be evaluated on the basis of the increase of the unexplained variation of the data from the expected value.

Multidimensional Scaling

Multidimensional Scaling (MDS) is a means for looking at the structure underlying a large number of associated measures. This, unlike factor analysis, does not require that relationships between the variables be explained based on correlations from a linear association. Multidimensional Scaling can, in fact be based on an interval scale, but it can also be also based on a rank ordering of the distances between various stimuli. In addition to MDS, there are various clustering techniques based on measures and methodologies that are not the traditional interval, continuous, and normally distributed measures.

Resampling Processes

"An important theme of what follows (resampling plans) is the substitution of computational power for theoretical analysis. The payoff, of course, is freedom from the constraints of traditional parametric theory, with its over reliance on a small set of standard models for which theoretical solutions are available. In the long run, understanding the limitations of the nonparametric approach should make clearer the virtues of parametric theory, and perhaps suggest useful compromises" (Efron, 1982, p. 3).

One of the newer developments in nonparametric statistics has been the creation of several methodologies that involve resampling subsets of the data. The most popular of these are Jackknifing and Bootstrapping. These methodologies were popularized by Efron in the late 1970s. Their purpose is to estimate the dispersion of various statistics from a set of data without requiring any assumptions about the distributions underlying a sample of data. Bootstrapping involves the random sampling of a large number of sub-samples from the original data set with replacement. Each sub-sample

contains the same number of elements as the original data set but does not have to contain all the original elements of the data set. By resampling a large number of times and producing the statistic at each iteration, the distribution of the statistic can be obtained. In Jackknifing there is a limited number of data sets created, each of which contains original data but omits at least one of the original data elements. Jackknifing requires far fewer computations but Bootstrapping is normally considered to be superior. (Efron & Gong, 1983) Various programs can be used in these resampling strategies including SPSS with some modifications to the basic package.

Other Considerations when Using Nonparametric Statistical Tools

About the Central Limit Theorem and Law of Large Numbers

This theorem and law support the use of parametric statistics and normal approximations to probabilities associated with nonparametric tests of location when there is an interval scale but the population distribution can not be assumed to be normal. "The central limit theorem tells us that a sampling distribution always has significantly less wildness than the population it's drawn from. Additionally, the sampling distribution will act more and more like a normal distribution as the sample size is increased, **even when the population itself is not normally distributed! ...The law of large numbers** is even more basic than the central limit theorem and so can be considered more important. It says essentially that probability and statistics can only predict overall results for a large number of data points or trials....think "central limit theorem" when changing the sample **size** and think "law of large numbers" when changing the **number** of samples" (Rogers et al., 2004).

About Ordinal Data and Ties

In the analysis of ordinal data, one of the issues is how to deal with tied observations. In many of the statistics, the assumption is made that the underlying measure is a continuous measure. This removes the likelihood of ties because the probability that a specific point occurs twice is zero. In reality the specific score is to some level of precision and this causes ties in the various cases. In cases where the measure is the comparison of pairs of observations such as in various pairs tests, the likelihood of a tie is smaller, especially if the rater is given options of "larger" or "smaller." Gibbons (1971, p. 96-97) identifies several ways to deal with ties. Five approaches are described below:

- *Average ranking:* One of the simpler methods for dealing with tied observations is to use the average of the ranks for which the observations would have been tied. This maintains the average or sum of the ranks. It also, however, reduces the variance of the ranks.

- *Remove the observation:* This seems to be a methodology for the comparison of pairs of observations where the intent is to assign a binomial outcome to the results of the comparison such as in the sign test or the median test where an observation is tied with the median.

- *Develop adjustments to the statistic:* This is often done for statistics based on the concordant or discordant pairs. For example, Kendall's Tau (τ) has one form that takes ties into account by reducing the number of pairs to the number of untied pairs (tau-b). Another example is the correction for ties that has been developed for the Kruskal-Wallis Test (Agresti, 1984) although some simply suggest that ties across groups can be resolved with average ranks or by using an assignment that gives the lowest chance to reject the null hypothesis (Gibbons, 1971).

- *Calculate all alternatives:* Assign ranks to tied observations in all possible ways and compute the statistic under the various alternatives. Then one can either use the average of the statistics or use the most conservative of the statistics. This is similar to the assignment of tied ranks that seems to produce a conservative estimate of the true difference.

- *Select ranks at random before the experiment.* This seems to preserve the random character of the situation. If one had randomly assigned ranks and two observations were tied, one would use the random ranks to break the tie. If the random items were a sequence, then one would use the random assigned sign to the item tied with the median.

If there are a large number of ties, one may need to do additional analysis or pre-process the data. Also, be sure there is an understanding of how SPSS handles tied observations and/or ranks.

Acknowledgement

The authors thank Dr. Peter Ammerman for his major contributions to an earlier version of this chapter.

References

Conover, W. J. (1971). *Practical nonparametric statistics.* New York, NY: John Wiley and Sons, Inc.

Efron, B. (1982). The jackknife, the bootstrap and other resampling plans, Philadelphia, PA: *Society for Industrial and Applied Mathematics.*

Efron, B. & Gong G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician* 37, 36-48.

Ender, P. (2004). *Linear statistical models.* At website Education 2 30 BC, UCLA Department of Education, http://www.gseis.ucla.edu/courses/ed230bc1/notes1/cor3.html, Retrieved December, 2004.

Gay, L. R. & Airasian, P. (2003). *Educational Research: Competencies for Analysis and Applications.* (7th ed). Upper Saddle River, NJ: Merrill Prentice Hall, Pearson Education Inc.

Gibbons, J. D. (1971). *Nonparametric Statistical Inference.* New York, NY: McGraw-Hill.

Gravetter, F. J. & Wallnau, L. B. (2004) *Statistics for the behavior sciences* (6th ed). Belmont, CA: Wadsworth/Thomson Learning Inc.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1998). *Applied statistics for the behavioral sciences* (4th ed). Boston, MA: Houghton Mifflin Company.

Mertler, C. A. & Vannatta, R. A. (2002). *Advanced and multivariate statistical methods* (2nd ed). Los Angeles, CA: Pyrczak Publishing.

Rogers, T., Rogers, S. R., & Rogers, M. (2004). The central limit theorem — How to tame wild populations, At website Intuitor, http://www.intuitor.com/statistics/CentralLim.html, Retrieved November, 2004.

Sheskin, D. J. (1997). *The handbook of parametric and nonparametric statistical procedures.* New York, NY: CRC press.

Thierauf, R. J. (1997). *On-line analytical processing systems for business.* Westport, CT: Quorum Books.

Zar, J. H. (1984). *Biostatistical analysis* (2nd ed). Englewood Cliffs, NJ: Prentice-Hall.

Chapter 2
Analysis of Variance Applications in Institutional Research

Robert J. Ploutz-Snyder

This chapter will provide the institutional researcher with an overview of the basic theory behind the Independent and Repeated Measures Analysis of Variance (ANOVA) statistical techniques. The ANOVA statistic is a very versatile tool that can be put to task to address numerous important institutional questions, and if we are successful in this chapter, the reader will come away with several concrete applications of this statistical technique.

We will begin with a discussion of what types of data are appropriate for ANOVA, and how to collect and organize data for analysis. Because ANOVA is so versatile, we will then discuss some of the basics of statistical/ experimental design concepts that the analyst must be knowledgeable about in order to maximize the potential knowledge-gain from a study utilizing ANOVA. We will then discuss how one might use SPSS software to run the ANOVA statistic, including several specific examples in institutional research that illustrate the different experimental designs that ANOVA can handle. Some of the data sets used in this chapter were created specifically for instructional purposes, while others are actual data sets collected and analyzed with ANOVA. The end of this chapter will include some pointers on how the institutional researcher can use graphical techniques and statistically jargon-free language to present the results of a study analyzed with ANOVA to a statistically naive audience. Thus, our six learning objectives are to enable the reader to (1) Determine when it is appropriate to apply ANOVA; (2) Design data collection and management strategies appropriate for ANOVA; (3) Establish appropriate hypotheses for ANOVA; (4) Use SPSS to run various ANOVA models; (5) Interpret SPSS results from ANOVA runs; and (6) Report the findings of ANOVA research to a statistically naive audience.

Before we begin, I would like to acknowledge three of my many statistical mentors who taught me to use and love ANOVA during my graduate training. Dr. Allan J. Nash, Professor Emeritus at Florida Atlantic University, devoted tremendous hours and dedication to teaching statistics and experimental design, and I learned much of what I know about ANOVA from him. I also owe a debt of gratitude to Professors Francis S. Bellezza and Bruce W. Carlson at Ohio University, for further expanding my knowledge in multivariate and log linear analysis of variance during my doctoral training there. While this chapter serves as a hands-on ANOVA primer for the institutional researcher, it is certainly no surrogate for the in-depth training that one receives from semester-long courses taught by dedicated college and university faculty such as these.

Statistical & Theoretical Background for ANOVA

Like all statistics, ANOVA has assumptions about the data to be analyzed. ANOVA has been shown to be robust to modest violations of these assumptions, but if your data show considerable violations of one or more assumptions, you should consider data transformations or other statistical options that correct for serious violations of these assumptions. I will discuss some of these options in our examples later in the chapter.

The first assumption of the ANOVA statistic is that the outcome is collected randomly from the population with equal or similarly sized samples per group. This assumption is one that the researcher can usually control at the study design and data acquisition stages. If done well, it greatly increases the probability that the other assumptions will be met.

The second assumption of the ANOVA statistic is that the outcome of interest is measured on an interval scale, and is normally distributed. Sometimes it's necessary to transform data in order to meet this assumption, but ANOVA has also been shown to be robust to violations of this assumption (i.e. should perform adequately in the face of moderately non-normal data). Note that ANOVA assumes normality for the entire sample and within groups ("multivariate normality"). Institutional Research and Assessment examples of data commonly analyzed by ANOVA include age, salary, and GPA. Technically, Likert-scaled survey items are not interval in scale; however, these types of data are commonly evaluated with ANOVA throughout the literatures of education, psychology, assessment, medicine, and others. However, one should note that these studies usually combine Likert-scaled survey items that assess similar constructs together by calculating the sum or average of several Likert-scaled items in which case Likert-scaled survey data tends to meet the assumptions of ANOVA. It is rare that a single Likert-scaled survey item would meet the assumptions of ANOVA, and thus should probably be analyzed using non-parametric statistical techniques.

Third, ANOVA assumes that the variance on the outcome variable is similar across groups. This assumption is referred to as "homogeneity of variance." It is something we can test for and correct with SPSS, though ANOVA has been shown to be robust to modest violations to this assumption. There is also a special kind of "homogeneity of variance" for repeated-measures ANOVA designs, called "sphericity." Sphericity is when the variance of the *difference* between the estimated means for any pair of groups is the same as for any other pair. It is also something we can test for and correct with SPSS if necessary.

Weiner, Brown and Michels (1991) present a more in-depth discussion of the assumptions of the ANOVA statistic, and how to test for them statistically. Tabachnick and Fidell, (1989) devote an entire chapter to this subject, including a nice discussion of the effect of data transformations on meeting ANOVA assumptions.

A question that arises when discussing ANOVA is, "Why use ANOVA,

when there's a perfectly good t-test in my toolbox?" In the first monograph on statistics and institutional research, published by Coughlin and Pagano (1997), one reads about how researchers can compare continuous-level data, such as students' GPA, across two groups, (for example, males versus females, or School of 'X' versus School of 'Y') using either an Independent Measures t-test (if you are comparing two different groups), or a Repeated Measures t-test (if one is comparing the same subjects measured two different times). T-tests are taught the world over as part of an "intro stats" course in all disciplines, and the intuitive nature of the test makes it one that practically everyone understands and accepts. So why not use the t-test in place of ANOVA?

The answer has to do with situations in which there are more than two groups to compare. Institutional researchers are often interested, for example, in comparing student-learning outcomes across several majors, departments, or schools. Admissions people might like to compare the incoming qualifications of entering students by high school and the many high schools applicants hail from. Perhaps one would like to compare alumni donations received by college major or year of graduation. What about comparing student satisfaction ratings across various residence halls, food service facilities, or ratings of faculty teaching across departments? One could argue that running as many multiple t-tests as necessary, each comparing pairs of departments, faculty, majors, high schools, etc., would indicate where pair-wise differences by group exist. Yet there is a large flaw in this logic, for every time we run a t-test comparing two groups (or the same group at two times), we risk making a Type-I error. Remember that a Type-I error is one in which we falsely reject the null hypothesis—claiming that there is a statistical difference between groups when the difference observed is purely due to chance. While the risk of making this sort of error for a single t-test is very low, determined by alpha (usually .05), the trouble with running *multiple* t-tests making all pair-wise comparisons is that the experiment-wise risk increases at the rate of alpha per additional test. For example, if we wanted to compare student ratings of faculty teaching across ten different departments to determine which departments were rated significantly higher than others, this would require 100 different pairwise comparisons (A vs. B, A vs. C, A vs. D....A vs. J, B vs. C, ...), each contributing .05 alpha risk. During the course of this study, we would have .05*100, or alpha = 5.0 risk of incorrectly concluding that there was a significant difference in student ratings of faculty! Clearly that is an unacceptable statistical risk.

ANOVA in its simplest form (one-way ANOVA) is appropriate for comparing more than two groups, or one group measured more than two times. Within this chapter, we will illustrate how to use ANOVA to determine whether or not there is an overall difference between groups (or repeated observations), and if so, then how to determine where pair-wise differences exist if that is of interest.

ANOVA is also useful for evaluating the effects of more than one grouping factor on an outcome. For example, if we wanted to compare faculty salary by department *and* gender, or if we wanted to research the effects of a human resources policy change on staff satisfaction before versus after the change, *and* by administrative level, ANOVA is a flexible analytic tool that can accommodate two, three, four or more factors simultaneously. In fact, there is no limit to the number of factors ANOVA can use, though I will discuss practical limitations that I hope will moderate some potentially overzealous experimental designs.

Two General Types of ANOVA: Independent vs. Repeated Measures Designs

Similar to the t-test, ANOVA can handle independent and repeated measures designs. Studies that compare two or more separate, independent groups, like males versus females, or full versus associate versus assistant professor, are known as independent-measures designs, and will be analyzed using an Independent Measures ANOVA (IM-ANOVA). Studies that measure the same outcome from the same sample, but on two or more occasions (pre/post studies, or freshmen, sophomore, junior, senior longitudinal studies) are commonly referred to as longitudinal or repeated measures designs, and will be analyzed using the Repeated Measures ANOVA (RM-ANOVA) technique. Unlike the t-test, ANOVA can also handle mixed-model designs that assess the effects of one or more of each type of factor. These are commonly referred to as mixed-model ANOVA designs.

We will discuss and illustrate the IM-ANOVA first, followed by RM-ANOVA, ending with an example of a mixed-model ANOVA.

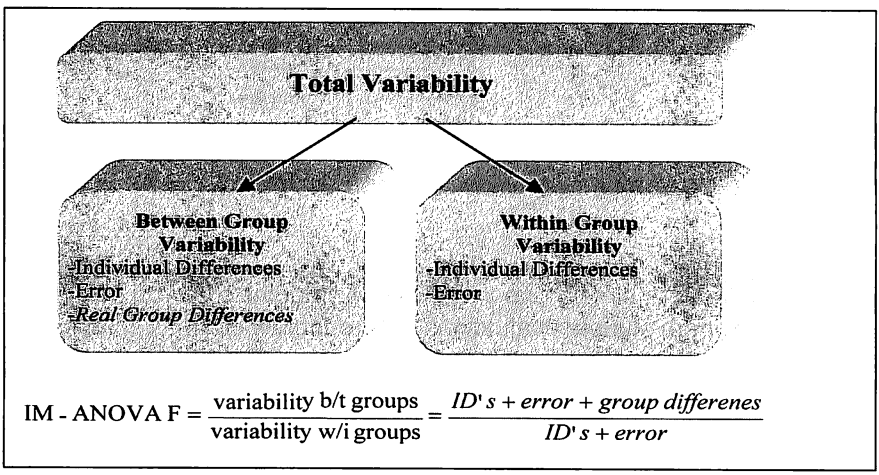
One-Way ANOVA: The Simplest Independent Measures ANOVA Design

The simplest of research questions appropriate for ANOVA would be a comparison of three or more independent groups[1], with the null hypothesis that no differences exist across groups, and the alternative hypothesis that there is an overall group difference. For illustrative purposes, suppose there are k=three independent groups in the population that one would like to make inferences about, and that one randomly selected equal-sized samples of n=6 from each population, resulting in a total of n=18 subjects. The name, "analysis of variance," refers to the fact that the ANOVA statistic breaks down the total variability (i.e. variance) in the n=18 subjects into two major components—"between group" variability and "within group" variability. In so doing, ANOVA then determines whether any "between group" variability is unusual, given the "within group" variability. In other words, are the subjects *more different* between groups than they are within groups?

Variability within groups is comprised of individual differences among respondents within that group (individual differences), plus error variance. Variability between groups is due to any real group differences that exist,

plus individual differences and error variance. The ANOVA statistic evaluates the proportion of total variability that is due to between-group differences, and if there is an unusual amount of between-group variability, the statistic is said to be significant, and one would reject the null hypothesis of no differences among groups. Figure 1 illustrates the theory behind the ANOVA statistic.

Figure 1
Components of the IM-Analysis of Variance F-Statistic



Note that if the data meet the homogeneity of variance assumption for ANOVA, by simple algebraic rules, we know that the individual differences and error components in the numerator of the F-ratio will "cancel out" those terms in the denominator of the ratio, leaving nothing but group differences in the numerator over 1 in the denominator. Thus the F-ratio is a quantity that represents the amount of variance due to between-group differences. If there are no differences, then F=1 because the numerator and denominator of the fraction will be equal. As the between-group differences increase, so too does F. The ANOVA statistic (F) is derived from a known distribution, called the F-distribution, and like the t-table for statistical significance, there is a F-table that associates probability values to observed F-values, given sample size, and degrees of freedom from the numerator and denominator of the F-ratio. The F-table is commonly available in most statistical texts, though modern statistical software provides *p*-values associated with observed F-ratios, given the known properties of the F-distribution.

As mentioned previously, ANOVA analyzes variability. ANOVA uses the sum of squares (SS) as the raw measure of variability, adjusted by the appropriate degrees of freedom given the number of observations in the sample, and the number of groups (k) that are being compared. The adjusted SS terms are called mean squares (MS) by ANOVA nomenclature. All of these

components of the ANOVA F-statistic are summarized by modern statistical software in an ANOVA summary table. The ANOVA summary table shows the sum of squares, degrees of freedom, and mean squares attributable to between group differences, within group differences, and total. Table 1 shows the components of an ANOVA summary table.

Table 1
Components of the IM-ANOVA Summary Table

Source	df	SS	MS	F	*p*
Between Groups	k-1	$\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{G})^2$	$\frac{SS_{\text{between groups}}}{df_{\text{between groups}}}$	$\frac{MS_{\text{between groups}}}{MS_{\text{within groups}}}$	Probability given α, df_{between} and df_{within} of no differences
Within Groups	n-k	$\sum_{i=1}^{k} SS_{\text{w/i each group}}$	$\frac{SS_{\text{within groups}}}{df_{\text{within groups}}}$		
Total	n-1	$\sum_{i=1}^{n} x^2 - \frac{(\sum x)^2}{n}$			

To illustrate a simple one-way ANOVA, imagine that you have been asked to compare student evaluations of introductory courses across three departments: psychology, biology, and chemistry. For this example, let's assume that you randomly selected evaluations from $n = 253$ students, split nearly equally across the three departments. After running the One-way ANOVA command within the Compare Means menu on SPSS, and choosing Descriptive Statistics, Homogeneity Tests, and Tukey's HSD Post-Hoc options, SPSS produced the output shown in Figure 2. Note that our sample sizes were approximately equal per group, with n = 86, 78, & 89 student evaluations from psychology, biology and chemistry, respectively. Mean (SD) evaluations are presented in the descriptive statistics table, followed by the Levene's test for homogeneity of variance. The Levene statistic tests the null-hypothesis that the variance among groups is constant. In this case, the Levene's test was not significant, indicating that our data do not violate ANOVA's homogeneity of variance assumption. The ANOVA summary table shows an $F = 11.89$, with $p < .001$. This is clearly a significant result at alpha = .05, .01, or .001, indicating a significant overall difference in student evaluations of introductory courses across the three departments of psychology, biology and chemistry. This is known as the "omnibus F-test" because it tells us whether or not there is an *overall* difference among means from these three groups. However, the omnibus F-test does *not* tell us anything about potential differences between any two departments. For one to establish whether or not any two departments differ significantly in this study, it is necessary to run a post-hoc analysis evaluating the pair-wise comparisons. Several different choices of post-hoc tests are available for a researcher to choose; each designed to test for pair-wise differences, and each with a distinct set of assumptions. Here, I choose the Tukey's HSD post-hoc analysis

Figure 2
Results of a One-Way ANOVA Comparing Student Evaluations of Introductory Freshman Courses in Psychology, Biology, and Chemistry

Descriptives

Student eval of Intro course

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Psych 101	86	6.65	1.713	.185	6.28	7.02	4	9
Bio 101	78	5.49	1.778	.201	5.09	5.89	3	8
Chem 101	89	5.53	1.803	.191	5.15	5.91	3	8
Total	253	5.90	1.840	.116	5.67	6.13	3	9

Test of Homogeneity of Variances

Student eval of Intro course

Levene Statistic	df1	df2	Sig.
.164	2	250	.849

ANOVA

Student eval of Intro course

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	74.126	2	37.063	11.891	.000
Within Groups	779.202	250	3.117		
Total	853.328	252			

Multiple Comparisons

Dependent Variable: Student eval of Intro course

Tukey HSD

(I) Department	(J) Department	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Psych 101	Bio 101	1.16	.276	.000	.51	1.81
	Chem 101	1.12	.267	.000	.49	1.75
Bio 101	Psych 101	-1.16	.276	.000	-1.81	-.51
	Chem 101	-.04	.274	.988	-.69	.60
Chem 101	Psych 101	-1.12	.267	.000	-1.75	-.49
	Bio 101	.04	.274	.988	-.60	.69

* The mean difference is significant at the .05 level.

because Tukey's HSD test is one of the more conservative (and thus popular) options that hold experiment-wise error to .05. Reading the Tukey's HSD table, I can see that the mean difference in ratings of Psych 101 and Bio 101 is 1.16 (SE .276), with a *p*-value <.001. Student ratings of Psych 101 courses were also significantly higher than Chem 101 ($p < .001$). However, student evaluations of Bio 101 were not significantly different from evaluations of Chem 101, as noted by the mean difference score of .04 (SE = .274) and non-significant *p*-value of .989.

In this example, the researcher had no apriori hypothesis regarding which pairs of departments might differ from one another, and thus a simple one-way ANOVA followed by a post-hoc comparison of all possible pairs is appropriate. The choice of which post-hoc comparison to run can be involved, though the most common post-hoc analyses in the social/education literature is probably Tukey's HSD, because it adjusts for multiple comparisons appropriately without being overly conservative.

Bonferonni, Scheffe, and Sidak are other common post-hoc analytic choices that adjust for the multiple comparisons in slightly different ways. The Bonferonni correction is a simple, conservative way for adjusting for inflated risk of making Type I errors. When running multiple statistical analyses, the Bonferonni technique simply divides the critical alpha (typically .05) by the number of comparisons conducted before concluding that a significant effect is observed in order to maintain the experiment-wise Type I error risk. The Bonferonni adjustment is a more conservative adjustment technique than the others, and is thus preferred when the consequences of making a Type I error are severe. The Sidak and Scheffe are a little less conservative than the Bonferonni adjustment, and are sometimes preferred over Bonferonni. Note also that SPSS provides several other post-hoc choices, including the LSD (least squared difference) statistic that performs uncorrected multiple comparisons, and is thus a more liberal approach to post-hoc analysis. I generally recommend against using the LSD method in most situations. Finally, note also that in our case, the data met the assumption of homogeneity of variance. In situations in which the data violates this assumption, there are post-hoc tests available that attempt to correct for the heterogeneity of variance across groups when making the pair-wise comparisons (ex. Dunnett's, or Games-Howell). These tests also vary slightly in how they protect against Type I error rates, though the differences tend to be subtle.

We will discuss an alternative approach to running post-hoc pairwise comparisons (apriori contrasts) in the Repeated Measures ANOVA (RM-ANOVA) section.

Two-Factor Independent Measures ANOVA Designs

The ANOVA statistic in the more general form expands on the notion of splitting variance into between groups and within groups variance for more than one factor, and assessing the extent to which between-group variance is significant. ANOVA can, for example, determine whether data are significantly different between groups of students with different national origins, *and* by their choice of college major. The experimental design and analysis becomes more complicated, but is often far more interesting than simply a "one-way ANOVA times two."

In two-factor ANOVA designs, we actually assess the significance of *three* factors that contribute variance to the model. Obviously we evaluate

the effects of each of the two factors in the model (ex. national origin, college major). The additional effect that a two-factor ANOVA design evaluates is called the *interaction term,* and it represents the combined effects of both main effect factors. Interaction effects are often very interesting, seldom explored under multiple regression or other analytic approaches, and seem to me to represent the hallmark feature of multi-factorial ANOVA that makes this statistic so incredibly powerful and interesting. I will illustrate with another example.

In this example, you are examining students' first-term college GPA by college major, a three-level factor (Math, Business, U.S. History) *and* students' national country of origin/citizenship (assume here that students remain citizens of their country of origin). To simplify matters, assume that we have collapsed students into two groups on this latter factor; U.S. and non-U.S. citizens. Figure 3 shows the descriptive statistics and the results of the two-factor independent measures ANOVA. Note also that these data violate ANOVA's assumption of equal variance across groups indicated by a significant Levene's test for equal variance. While this is an undesirable reality, statistical studies have previously shown that ANOVA tends to be robust to this sort of problem, and while it remains necessary for the researcher to acknowledge this violation in any formal report or publication, the results of an ANOVA that violates the homogeneity of variance assumption tend to be accurate nonetheless.

What is clearly obvious when comparing Figure 3 to Figure 2, is that the ANOVA summary table in Figure 3 is much larger than the ANOVA table in Figure 2. We now have an analysis that evaluates five different F-ratios and associated *p*-values. The "corrected model" F-ratio is an evaluation of how well the model "fit the data," and is analogous to the evaluation of a multiple regression model. The "intercept" term is also analogous to a multiple regression model, as it represents the Y-intercept term for a linear model predicting GPA. In ANOVA, these two effects are not meaningful. For our purposes, only the remaining three effects (illustrated in bold-face font) are relevant—two main effects, and one interaction term. Specifically, we will evaluate whether students' GPA differs by Major (a main effect with 3 levels), and/or by Citizenship (a main effect with 2 levels), and/or whether or not there is a difference in college GPA that is dependent on *both* major *and* citizenship (a two-factor interaction). Each of these effects is represented in the ANOVA summary table as a separate row with their respective F-ratios and associated *p*-values.

Working through the main effects first, we see a significant effect for college Major ($p = .003$), and a non-significant effect for Citizenship ($p = .212$). Judging by this information alone, we might conclude that students' first-term college GPA differs by Major, but there's no difference based on their Citizenship. We might then further examine the mean GPAs across the three college majors in this study to determine which majors are significantly

Figure 3
Results of a 2 (citizenship) X 3 (major) IM-ANOVA Evaluating First-Term College GPA

Descriptive Statistics

Dependent Variable: GPA after 1 term

MAJOR	Citizenship	Mean	Std. Deviation	N
Math	U.S.	2.9750	.29580	20
	Other	3.5250	.29580	20
	Total	3.2500	.40351	40
Business	U.S.	2.6384	.84313	20
	Other	2.8240	1.00387	20
	Total	2.7312	.91984	40
US History	U.S.	3.5813	.31154	20
	Other	2.3965	.75263	20
	Total	2.9889	.82656	40
Total	U.S.	3.0649	.66571	60
	Other	2.9152	.86902	60
	Total	2.9900	.77447	120

Levene's Test of Equality of Error Variances[a]

Dependent Variable: GPA after 1 term

F	df1	df2	Sig.
16.165	5	114	.000

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept+MAJOR+CIT+MAJOR * CIT

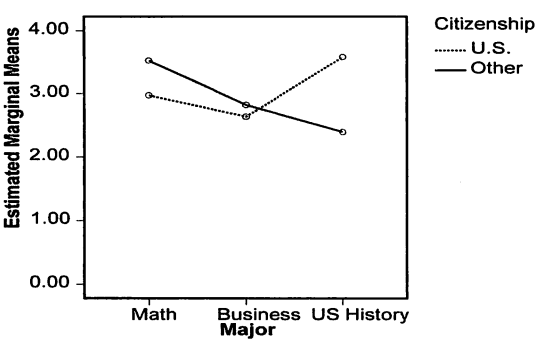
Tests of Between-Subjects Effects

Dependent Variable: GPA after 1 term

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	22.791[a]	5	4.558	10.695	.000
Intercept	1072.835	1	1072.835	2517.284	.000
MAJOR	**5.384**	**2**	**2.692**	**6.316**	**.003**
CIT	**.673**	**1**	**.673**	**1.578**	**.212**
MAJOR * CIT	**16.735**	**2**	**8.367**	**19.633**	**.000**
Error	48.585	114	.426		
Total	1144.211	120			
Corrected Total	71.376	119			

a. R Squared = .319 (Adjusted R Squared = .289)

Estimated Marginal Means of GPA after 1 term



different from the others, using post-hoc pair-wise comparisons. Regardless of what the pair-wise comparisons reveal, this approach ignores the potential impact of Citizenship to GPA, and in this example, where we have a highly significant Major by Citizenship interaction term (p = .000), we would be missing some very important information about the determinants of students' first-term GPA. The main effect that we see for Major is said to be *qualified by* the significant Major by Citizenship interaction term, and thus is it important that we consider the interaction term in our model *first,* before we move to an interpretation of any significant main effects revealed.

The interpretation of interaction terms, especially those involving more than two factors, can get complicated. I find it easier to visualize the effects of multiple factors by using the graphing features of SPSS (or other) statistical software (note the line-chart imbedded in Figure 2). An appropriate statement given our significant Major x Citizenship interaction term would be that the differences in GPA across college major that we see for U.S. citizens are not the same as the differences we see across major for non-U.S. citizens. From here, there are two general schools of thought as to what to do next to further understand the effects. One school of thought is to simply stop analyzing the data and interpret what is seen. Statisticians subscribing to this paradigm argue that the significance of the interaction term is sufficient enough information to halt further analyses, and that further analysis merely increases the experiment-wise risk of making a Type I error (falsely concluding that a difference exists, when it is due to chance alone).

A second school of thought is that the significant interaction term *gives us* the justification we need to break down the analysis further and determine more precisely how differences across one factor involved in the interaction compare across the levels of the other factor. That is, how do GPA differences between Majors for U.S. Citizens compare to GPA differences between Majors for non-U.S. Citizens? I generally subscribe to the latter school of thought in that a significant interaction term provides justification for follow-up analyses, even though we are technically increasing Type I alpha risk by conducting additional analyses on the data. If I am particularly concerned about this additional risk, I could employ a statistical correction for multiple analyses, such as the Bonferonni alpha-correction.

If we subscribe to the notion that additional analyses are warranted, we are not finished with this example just yet. Our next choice regards which factor we want to hold constant while examining differences across the levels of the other factor. We could, for example, look at the GPA differences between U.S. and non-U.S. Citizens within levels of Major. Or, we could examine differences across Major first for U.S. Citizens, then for non-U.S. Citizens. This choice is not trivial, and depends largely on the context in which one sets out to conduct this research in the first place. It would not be appropriate to try it both ways, as this further increases our probability for making Type I errors, and statisticians universally agree that such an unguided

approach (i.e. fishing expedition) is inappropriate. Let theory be the guide instead, and even then consider an alpha adjustment, like the aforementioned Bonferonni corrections. In this example, I'm interested in determining whether student GPA differs across major within level of citizenship. I'm less interested in whether U.S. and non-U.S. students differ in GPA within a specific college major. Therefore, I will hold the level of Citizenship constant and examine effects that college Major has on student GPA. Because our data include three majors, I will also conduct post-hoc pair-wise comparisons. As it turns out, the data in this example fail to meet the assumption of homogeneity of variance, and the Games-Howell post-hoc procedure makes adjustments for this fairly common situation, thus I choose the Games-Howell post-hoc procedure.

There are competing schools of thought as to the most appropriate way to conduct these follow-up analyses when an ANOVA interaction term is significant. The more common approach that we see in the literature is to conduct "simple-effects" analyses, and we will present this technique in the pages to follow. Before we get to that however, I would like to present a simpler alternative approach that clearly illustrates what the "simple-effects" analysis does, with some caveats.

Remember that the "big picture" here is that our initial ANOVA revealed a significant interaction of Major*Citizenship, meaning that the GPA difference by Major for U.S. citizens, are not the same as they are for non-U.S. citizens. One way to clarify these effects would be to restrict our sample to only U.S. citizens, and explore the GPA differences among the three college Majors, then do the same thing for non-U.S. majors. The SPSS' "split-file" command makes it particularly easy to conduct this sort of follow-up analyses, where we are literally conducting two separate one-way ANOVAs—one per each level of Citizenship. Figure 4a shows the abbreviated results of the two follow-up one-way ANOVAs, with Split-File set to compare the results of U.S. versus non-U.S. students. That is, the components of the ANOVA output for U.S. citizens are combined together with the non-U.S. citizens ANOVA output so that it is easier to compare the effects of Major—precisely what we want to do in order to better understand our interaction effect. In this case, the ANOVA summary tables show significant GPA differences across Major for both U.S. ($F = 15.30$, $p = .000$) and non-U.S. ($F = 11.72$, $p = .000$) citizens. However, the interesting and important results are in the post-hoc comparisons that detail the pair-wise GPA differences by Major. These pair-wise effects are illustrated in the line-chart that is imbedded in Figure 3. The data show that U.S. Citizens majoring in U.S. History have a significantly higher mean first-term GPA than U.S. Citizens majoring in either Business or History. On the contrary, non-U.S. Citizens majoring in Math show the highest first-term mean GPA, significantly higher than their non-U.S. Citizen counterparts majoring in Business or U.S. History.

The above "split-file" approach to understanding the significant interaction effects of major and citizenship on GPA provides an intuitive understanding of

Figure 4a
Follow-Up One-Way ANOVAs on Major, by Citizenship

Tests of Between-Subjects Effects

Dependent Variable: GPA after 1 term

Citizenship	Source	Type III Sum of Squares	df	Mean Square	F	Sig.
U.S.	Corrected Model	9.134[a]	2	4.567	15.301	.000
	Intercept	563.619	1	563.619	1888.324	.000
	MAJOR	**9.134**	**2**	**4.567**	**15.301**	**.000**
	Error	17.013	57	.298		
	Total	589.766	60			
	Corrected Total	26.147	59			
Other	Corrected Model	12.985[b]	2	6.492	11.721	.000
	Intercept	509.889	1	509.889	920.545	.000
	MAJOR	**12.985**	**2**	**6.492**	**11.721**	**.000**
	Error	31.572	57	.554		
	Total	554.446	60			
	Corrected Total	44.557	59			

a. R Squared = .349 (Adjusted R Squared = .326)

b. R Squared = .291 (Adjusted R Squared = .267)

Multiple Comparisons

Dependent Variable: GPA after 1 term
Games-Howell

Citizenship	(I) MAJOR	(J) MAJOR	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
U.S.	Math	Business	.3366	.19980	.232	-.1629	.8361
		US History	-.6063*	.09606	.000	-.8406	-.3720
	Business	Math	-.3366	.19980	.232	-.8361	.1629
		US History	-.9429*	.20099	.000	-1.4447	-.4411
	US History	Math	.6063*	.09606	.000	.3720	.8406
		Business	.9429*	.20099	.000	.4411	1.4447
Other	Math	Business	.7010*	.23401	.017	.1137	1.2884
		US History	1.1285*	.18082	.000	.6778	1.5792
	Business	Math	-.7010*	.23401	.017	-1.2884	-.1137
		US History	.4275	.28055	.292	-.2589	1.1139
	US History	Math	-1.1285*	.18082	.000	-1.5792	-.6778
		Business	-.4275	.28055	.292	-1.1139	.2589

Based on observed means.
*The mean difference is significant at the .05 level.

our data, and this approach can arguably suffice for publication in peer-reviewed journals. In this case, the split-file approach is particularly attractive because our data failed to meet the assumption of homogeneity of variance, and thus we were able to employ the Games-Howell post-hoc technique on our pair-wise comparisons, thus controlling for the unfortunate reality of heterogeneity of variance that so often plagues research data. Nevertheless, many would argue that this technique is flawed because we ran multiple analyses without correcting for the inflated Type-I error rates. Therefore, we will turn our attention to conducting the more popular "simple-effects" approach for follow-up

analyses of the same data, and then discuss some limitations to either approach.

Conducting a simple-effects follow-up analysis on our GPA data with SPSS requires use of the Syntax window, as the appropriate code is not available through their graphical user interface. The syntax required to execute a simple-effects analysis, and the relevant output is presented in Figure 4b. Note that the ANOVA summary table in Figure 4b is exactly the same as in Figure 4a—the omnibus F-test is identical. However, the Pairwise Comparisons table and the Univariate ANOVA tests in Figure 4b show slightly different significance values than the "split-file" illustration in Figure 4a.

Let us first focus on the Univariate Tests in Figure 4b, in comparison to the Tests of Between-Subjects Effects table in Figure 4a. Both the "split-file approach" (Figure 4a) and the "simple-effects" analysis (Figure 4b) are attempting to do the same thing—conduct a one-way ANOVA within level of citizenship. Note, however, that the degrees of freedom-error (df_e) in these techniques are different. In the "split-file" approach (Figure 4a), the df_e for U.S. and non-U.S. citizens analyses are both equal to 57 (n per sub-group, minus k number of groups),

Figure 4b
SPSS Simple-Effects Follow-Up ANOVA Syntax and Output on Major, by Citizenship

SPSS Syntax for Simple

```
UNIANOVA
  term1gpa BY major cit
  /METHOD = SSTYPE(3)
  /INTERCEPT = INCLUDE
  /PLOT = PROFILE( major*cit )
  /EMMEANS = TABLES(major*cit) COMPARE(major) ADJ(SIDAK)
  /CRITERIA = ALPHA(.05)
  /DESIGN = major cit major*cit .
```

Tests of Between-Subjects Effects

Dependent Variable: GPA after 1 term

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Mode	22.791 [a]	5	4.558	10.695	.000
Intercept	1072.835	1	1072.835	2517.284	.000
major	5.384	2	2.692	6.316	.003
cit	.673	1	.673	1.578	.212
major * cit	16.735	2	8.367	19.633	.000
Error	48.585	114	.426		
Total	1144.211	120			
Corrected Total	71.376	119			

a. R Squared = .319 (Adjusted R Squared = .289)

Pairwise Comparisons

Dependent Variable: GPA after 1 term

Citizenship	(I) major	(J) major	Mean Difference (I-J)	Std. Error	Sig. [a]	95% Confidence Interval for Difference [a]	
						Lower Bound	Upper Bound
U.S.	Math	Business	.337	.206	.285	-.164	.837
		US History	-.606*	.206	.012	-1.107	-.106
	Business	Math	-.337	.206	.285	-.837	.164
		US History	-.943*	.206	.000	-1.443	-.443
	US History	Math	.606*	.206	.012	.106	1.107
		Business	.943*	.206	.000	.443	1.443
Other	Math	Business	.701*	.206	.003	.201	1.201
		US History	1.129*	.206	.000	.628	1.629
	Business	Math	-.701*	.206	.003	-1.201	-.201
		US History	.427	.206	.117	-.073	.928
	US History	Math	-1.129*	.206	.000	-1.629	-.628
		Business	-.427	.206	.117	-.928	.073

Based on estimated marginal means

* The mean difference is significant at the .05 level.

a. Adjustment for multiple comparisons: Sidak.

Univariate Tests

Dependent Variable: GPA after 1 term

Citizenship		Sum of Squares	df	Mean Square	F	Sig.
U.S.	Contrast	9.134	2	4.567	10.716	.000
	Error	48.585	114	.426		
Other	Contrast	12.985	2	6.492	15.233	.000
	Error	48.585	114	.426		

Each F tests the simple effects of major within each level combination of the other effects shown. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

Note that the bold-face font in the syntax highlights the command necessary to run the appropriate simple-effects analysis using Sidak post-hoc comparisons.

whereas the df_e in the univariate F-tests of the "simple-effects" analyses (Figure 4b) are both equal to 114. This difference is because of the way in which these two approaches differ. The "split-file" approach discussed previously makes no attempt at adjusting for the multiple comparisons that we conducted; it merely repeats the one-way ANOVA command for either sub-sample. In contrast, the "simple-effects" analysis bases the Univariate F-tests on the marginal means reflected in the pairwise comparisons table (adjusted in this case with Sidak corrections). Without getting into too much detail, this technique is a method for analyzing simple effects of one (or more) factor(s) on an outcome, *within levels* of another factor, while adjusting the df_e accordingly in an attempt to correct for potential Type-I error risks. Because there are six total pair-wise comparisons made in this model (U.S. Math vs. U.S. Business, U.S. Math vs. U.S. History, U.S. Business vs. U.S. Math, Non-U.S. Math vs. Non-U.S. Business, Non-U.S. Math vs. Non-U.S. History, Non-U.S. Business vs. Non-U.S. Math), the df_e in the error terms of the simple-effects analysis = n-6 (in our case 120-6=114). In our case, the *p*-values of the Univariate F-tests in either of the two techniques are so exceedingly small ($p < .001$) that the difference between the two analytic approaches is not observable; both approaches show a significant overall difference in student GPA across the three Majors—and this is true for U.S. and Non-U.S. citizens. Note, however, that the "simple-effects" analysis is more conservative because it makes the adjustment for inflated Type-I error risk in its determination of df_e, and this is the primary reason why many prefer the "simple-effects" approach.

Next, let us now turn our attention to pairwise comparisons table revealed in Figure 4b from the simple-effects analysis, for this is where, in my opinion, the "simple-effects" analysis has some notable disadvantages over the "split-file" technique. The options for the type of post-hoc comparisons using the "simple-effects" analysis are quite limited; one must choose the Bonferonni-adjusted comparisons, the Sidak-adjusted comparisons, or the LSD (unadjusted) pairwise comparisons. As discussed previously, the Bonferonni adjustment is a conservative correction for multiple pairwise comparisons, followed in level of conservatism by the Sidak-adjustment, and finally the LSD comparisons, which are unadjusted pairwise comparisons. Figure 4b shows the results of the Sidak pairwise comparisons. Note that the *p*-values of the Sidak comparisons in Figure 4b vary slightly from the Games-Howell comparisons in Figure 4a. If we were employing the usual critical alpha=.05, our significance-decisions on either approach would be the same, but in situations in which differences among groups were not as dramatic as these data, the two procedures could be at odds with one another on the results.

In contrast to the "simple-effects" analysis, the "split-file" approach described earlier offers a wider variety in the types of post hoc choices. Choices include the Bonferonni, Sidak, and LSD options available to "simple-effects" analyses, but also including other popular post hoc tests (ex.Tukey's

HSD), and the Games-Howell technique that adjusts for heterogeneity of variance problems among the cells of a multi-factorial study. This, in my opinion, represents a disadvantage of employing the "simple-effects" analysis on multi-factorial data in which the assumption of homogeneity of variance has been violated. When employing the "simple-effects" analysis, one does not have the option to run pairwise contrasts that adjust for heterogeneity of variance, like the Games-Howell, Dunnett's, or others, and this is a common situation in many disciplines, including Institutional Research and Assessment. Another advantage of the "split-file" approach is that in higher-order factorial designs (those with more than two factors), the "split-file" approach allows us to evaluate lower-order interaction terms, whereas the "simple-effects" analysis does not.

A major disadvantage of the "split-file" approach is that it *fails to adjust* for inflated Type-I error risk in running the two subsequent one-way ANOVAs (one per level of our Citizenship factor), but it *has the advantage* of allowing us to correct for heterogeneity of variance among the levels of our other factor (choice of major). On the other hand, the "simple-effects" analysis *does adjust* for the inflated Type-I error risk, but it *does not make adjustments* for homogeneity of variance situations. Alas, the analyst is forced to weigh the pros/cons of the two approaches. Unfortunately, we cannot provide an answer to this dilemma that will satisfy all situations, although most statisticians lean towards the use of "simple-effects" analysis in favor of the Type-I error reduction, and then defend the notion that ANOVA has been shown to be robust to violations of homogeneity of variance in Monte Carlo studies. For this reason, we will conduct "simple-effects" analyses in future illustrations that warrant follow-up analyses.

Regardless of which approach we take, it is clear that had we not explored the effects of the significant Major by Citizenship interaction term in the beginning, when we noticed a significant main effect for Major, we would have missed valuable information in understanding the relative student performance attributable to their national origin.

Three-Factor Independent Measures ANOVA Designs

As illustrated above, the ANOVA statistic is capable of parsing out the effects of two factors (main effects) on a continuous outcome variable, and in addition to evaluating these main effects, ANOVA also provides an analysis of the interaction between them. Indeed, ANOVA can expand this concept to three, four, or even more factors. And with a little practice, the reader will quickly surmise that running these more complex ANOVA models is no more difficult than running a two-factor design. However, understanding these more complex ANOVA models can become an arduous task. For example, a study examining the effects of three independent factors; gender, race/ethnicity, and student level (freshmen, sophomore, junior, senior), would involve an assessment of seven potentially meaningful effects:

- Three main effects
  - Gender
  - Race/ethnicity
  - Student level (freshmen, sophomore, junior, senior)
- Three two-way interaction effects
  - Gender * Race/ethnicity
  - Gender * Student level
  - Race/ethnicity * Student level
- One Three-way interaction effect
  - Gender * Race/ethnicity * Student level

Our approach to such a design would be to first evaluate the significance of the three-way interaction term. This is the highest-order term in this model, and if significant, it would qualify any other significant effects in the model, so we must start here. A significant three-way interaction term is a very complex effect that could require subsequent analyses based on the same logic described in the two-factor ANOVA paragraphs above. We would first need to hold one factor constant, then run separate two-factor ANOVAs on the subsequent factors. Each of these two-factor designs would be evaluated separately, possibly requiring additional follow-up One-Way ANOVA statistics, until we have fully broken down the model to its elemental effects. If the three- way interaction term in the original model were not significant, then we would look to the three two-way interaction terms and evaluate each of them accordingly, with potential One-Way ANOVA follow-up analyses.

Let us work through an example of the simplest of three-factor ANOVA designs—one in which all three factors have only two levels. In this example, the analyst is asked to determine whether or not there is a gender bias in faculty salaries in a study that uses tenure status (tenured, not) and department (Women's Studies, Biology) as two additional factors that may affect salary. Thus, this is a 2 (gender) x 2 (tenure status) x 2 (Department: Biology, Women's Studies) completely independent measures factorial ANOVA design, using faculty salary as the outcome variable. Note that these data again are fictitious—generated to illustrate the procedure of conducting and interpreting a 3-factor ANOVA statistic. To simplify matters, I have chosen to compare only two departments.

Figure 5 shows the SPSS output of the 3-factor omnibus F-test on these faculty salary data. The ANOVA summary table reveals a significant three-way interaction involving gender, tenure status, and department ($F = 7.18$, $p = .009$). This significant interaction means that all three factors have an impact on faculty salary, and so to interpret faculty salary by any of the lower-order interaction terms or main effect terms would be potentially misleading because those effects fail to capture the entire picture of salary differences caused by gender, department, and tenure status. To better understand this complex interaction, it is necessary to a run simple-effects

Figure 5
Three-Factor Completely Independent Measures ANOVA on Faculty Salary

Descriptive Statistics

Dependent Variable: SALARY

SEX	TENURE	DEPT	Mean	Std.	N
Female	Untenured	Woman's Studies	35353.90	2928.690	10
		Biology	30405.30	2924.151	10
		Total	32879.60	3815.441	20
	Tenured	Woman's Studies	45405.40	2960.507	10
		Biology	47142.60	1392.477	10
		Total	46274.00	2421.631	20
	Total	Woman's Studies	40379.65	5899.330	20
		Biology	38773.95	8870.688	20
		Total	39576.80	7480.085	40
Male	Untenured	Woman's Studies	34680.30	2342.079	10
		Biology	38834.20	4568.788	10
		Total	36757.25	4126.339	20
	Tenured	Woman's Studies	43543.20	2165.393	10
		Biology	61705.30	3979.357	10
		Total	52624.25	9824.854	20
	Total	Woman's Studies	39111.75	5048.832	20
		Biology	50269.75	12451.629	20
		Total	44690.75	10948.775	40
Total	Untenured	Woman's Studies	35017.10	2603.962	20
		Biology	34619.75	5712.646	20
		Total	34818.43	4386.643	40
	Tenured	Woman's Studies	44474.30	2699.130	20
		Biology	54423.95	8014.230	20
		Total	49449.12	7760.362	40
	Total	Woman's Studies	39745.70	5457.619	40
		Biology	44521.85	12155.452	40
		Total	42133.78	9665.491	80

Levene's Test of Equality of Error Variances [a]

Dependent Variable: SALARY

F	df1	df2	Sig.
1.783	7	72	.104

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept+SEX+TENURE+DEPT+SEX * TENURE+SEX * DEPT+TENURE * DEPT+SEX * TENURE * DEPT

Tests of Between-Subjects Effects

Dependent Variable: SALARY

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	6707882432.750[a]	7	958268918.964	102.605	.000
Intercept	142020399660.1	1	142020399660	15206.659	.000
SEX	**523049692.050**	**1**	**523049692.050**	**56.005**	**.000**
TENURE	**4281147649.800**	**1**	**4281147649.8**	**458.399**	**.000**
DEPT	**456232176.450**	**1**	**456232176.450**	**48.850**	**.000**
SEX * TENURE	**30568753.800**	**1**	**30568753.800**	**3.273**	**.075**
SEX * DEPT	**814560188.450**	**1**	**814560188.450**	**87.218**	**.000**
TENURE * DEPT	**535302045.000**	**1**	**535302045.000**	**57.317**	**.000**
SEX * TENURE * DEPT	**67021927.200**	**1**	**67021927.200**	**7.176**	**.009**
Error	672433641.200	72	9339356.128		
Total	149400715734.0	80			
Corrected Total	7380316073.950	79			

a. R Squared = .909 (Adjusted R Squared = .900)

analysis and determine the simple effects on salary within levels of one of our factors. Again, the choice of which factor to hold constant is largely dependent on the initial research question. Because we were charged to look for "gender biases" across the other factors, it is probably not a good idea to hold gender constant and look for salary differences across departments or by tenure status. Instead, I choose to hold department constant and examine the simple effects of gender and tenure status upon faculty status within departments. Figure 6 shows the results of our simple-effects follow-up analysis, where we can more clearly see that the effects of gender and tenure status on faculty salaries are not mirrored in these two departments. The results of our simple-effects analysis of faculty salaries in the Women's Studies department are straightforward—there are no significant gender differences in salary for the untenured or tenured faculty members in this department.

However, our analysis of faculty salaries in the Biology department reveals a very different story. Here, we see that there is a significant gender difference in salaries for both tenured and untenured faculty. Men receive higher salaries than women regardless of tenure status. Note that the two line-charts (scale-equilibrated) on the bottom of Figure 6 illustrate clearly this three-way interaction effect. In the Women's Studies department, the lines representing male and female faculty are very close to each other and parallel, but the lines representing male/female faculty in the Biology department are quite far apart, with male mean salaries higher than female mean salaries in both tenured and untenured faculty. The next challenge, of course, would be to speculate as to why a gender bias of this sort exists in the Biology department, and what sort of institutional practices may be promoting such discriminatory practices.

By now the reader should be able to appreciate that running three, four, or higher-order factorial ANOVA designs is not a trivial undertaking. With enough time and concentration, it is possible to deeply explore such effects, but the real task is yet to come—when the need to explain such complex results to a potentially less statistically motivated and trained audience arises. I will end this chapter with some tips on how to make this task easier, though in general I tend to avoid three-factor ANOVA designs, and I draw the line at four factor designs entirely. It has been my experience that most research endeavors involving more than three factors can usually be reduced to fewer factors, or split into smaller, more manageable sub-studies, which, when taken together, represent a far better understanding of the issue than a single large study.

Repeated Measures ANOVA Designs

Thus far, all of the designs that we have discussed involved comparisons of independent groups of observations—males and females, U.S. and non-U.S. Citizens, tenured and untenured faculty members. However, the ANOVA is not restricted to comparing independent groups of subjects. ANOVA can

Figure 6
Follow-Up Two-Factor Simple-Effects ANOVAS on Faculty Salary by Tenure and Gender, within Department

Estimates

Dependent Variable: salary

dept	sex	tenure	Mean	Std. Error	95% Confidence Interval	
					Lower Bound	Upper Bound
Woman's Studies	Female	Untenured	35353.900	966.403	33427.410	37280.390
		Tenured	45405.400	966.403	43478.910	47331.890
	Male	Untenured	34680.300	966.403	32753.810	36606.790
		Tenured	43543.200	966.403	41616.710	45469.690
Biology	Female	Untenured	30405.300	966.403	28478.810	32331.790
		Tenured	47142.600	966.403	45216.110	49069.090
	Male	Untenured	38834.200	966.403	36907.710	40760.690
		Tenured	61705.300	966.403	59778.810	63631.790

Pairwise Comparisons

Dependent Variable: salary

dept	tenure	(I) sex	(J) sex	Mean Difference (I-J)	Std. Error	Sig. [a]	95% Confidence Interval for Difference [a]	
							Lower Bound	Upper Bound
Woman's Studies	Untenured	Female	Male	673.600	1366.701	.624	-2050.868	3398.068
		Male	Female	-673.600	1366.701	.624	-3398.068	2050.868
	Tenured	Female	Male	1862.200	1366.701	.177	-862.268	4586.668
		Male	Female	-1862.200	1366.701	.177	-4586.668	862.268
Biology	Untenured	Female	Male	-8428.900*	1366.701	.000	-11153.368	-5704.432
		Male	Female	8428.900*	1366.701	.000	5704.432	11153.368
	Tenured	Female	Male	-14562.700*	1366.701	.000	-17287.168	-11838.232
		Male	Female	14562.700*	1366.701	.000	11838.232	17287.168

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

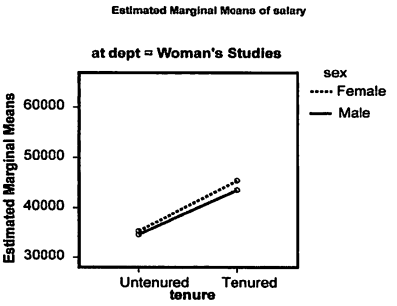
a. Adjustment for multiple comparisons: Sidak.

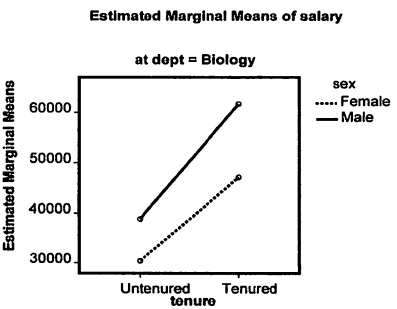
Univariate Tests

Dependent Variable: salary

dept	tenure		Sum of Squares	df	Mean Square	F	Sig.
Woman's Studies	Untenured	Contrast	2268685	1	2268684.800	.243	.624
		Error	6.72E+08	72	9339356.128		
	Tenured	Contrast	17338944	1	17338944.20	1.857	.177
		Error	6.72E+08	72	9339356.128		
Biology	Untenured	Contrast	3.55E+08	1	355231776.1	38.036	.000
		Error	6.72E+08	72	9339356.128		
	Tenured	Contrast	1.06E+09	1	1060361156	113.537	.000
		Error	6.72E+08	72	9339356.128		

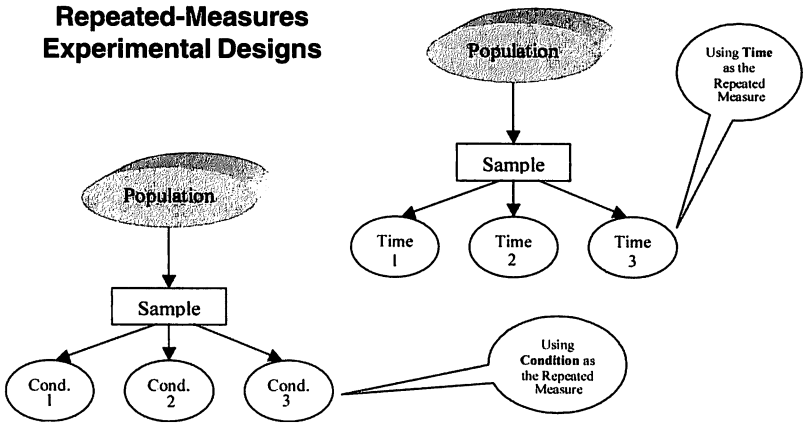
Each F tests the simple effects of sex within each level combination of the other effects shown. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.





also be put to the task of comparing repeated observations collected from the same individuals, commonly referred to as "pre/post" designs, or "repeated measures" designs. I will refer to these kinds of factors as repeated-measures designs (Figure 7).

Figure 7
Repeated-Measures Experimental Designs



An example of the simplest of repeated-measures designs might involve comparing student learning outcomes as measured by some reliable assessment instrument (e.g., "pre" versus "post" participation in a college program designed to improve student learning). With only two repeated observations (pre, post), such a simple design could easily be analyzed using the repeated-measures t-test, a modification of the independent-measures t-test that accounts for the correlation of repeated observations within subject.

However, if there is more than one repeated observation (ex. pre, post1, post2), then a more sophisticated ANOVA approach is necessary, just as we needed ANOVA to compare three or more independent groups in our discussion of the one-way IM-ANOVA. Repeated measures ANOVA is also useful when we want to examine the effects of more than one experimental condition (ex. different teaching modalities) on some outcome variable, controlling for potential differences among groups of individuals by using the same individuals across the various conditions (see Figure 7).

The general logic of the ANOVA approach for repeated-measures is very similar to IM-ANOVA, with one important distinction. Repeated-Measures ANOVA is evaluated by an F-ratio, where the numerator represents the differences across condition or time, and the denominator represents variation due the sample, plus error. A major difference in the calculations of the RM-ANOVA, however, is that because we are taking measurements from the *same sample* on *repeated occasions*, variability due to individual differences

are constant across time or condition, and thus variance attributable to individual differences is effectively removed from the model. Remember that because the Individual Differences (IDs) are components of variance in both the numerator and denominator, if the assumption of homogeneity of variance (sphericity) is met, they cancel each other out and are effectively removed from the model. The result is a statistical test that is more powerful than its independent-measures counterpart. This is an experimental design reality, and is a strength of repeated-measures statistical designs over independent-measures designs. This is not a property of the statistic; it is a property of all repeated-measures designs. Repeated-Measures designs (sometimes referred to as dependent-measures) are common in many literatures, including education literature, partly due to their intuitive appeal and simple application. "Pre/post" designs in which we measure something, intervene with an educational or other type of intervention, and then measure the same thing again—all with the same subjects—have a certain appeal. If we find a statistical difference pre-to-post, then it is fairly easy to accept the likelihood that our intervention was the cause of the change. One cannot argue that the observed difference is possibly due to the random effect of having a "better" sample in one group relative to the other, because we only *have* one group! Also, repeated measures designs require fewer subjects to detect the effect of an intervention—partly because we only need one sample, but also because the statistic is more powerful to begin with.

The data set preparation for a RM-ANOVA requires that the repeated observations be represented as separate columns in the data set, whereas before with IM-ANOVA, we had a single column representing our outcome variable, with grouping variables making up other columns. Also, because RM-ANOVA is technically a member of the "multivariate ANOVA" family of statistics, running a RM-ANOVA on modern statistical software, such as SPSS, usually requires commands executed from different menu options or syntax structure. Nevertheless, interpreting and understanding RM-ANOVA results is only subtly different from that of the IM-ANOVA, and the reader should be pleasantly surprised at how simple this task is, given the knowledge gained thus far.

Using Single-Factor Repeated-Measures ANOVA Where Time is the RM-Factor

We began our illustrations of IM-ANOVA with the simplest of designs, called the "one-way" ANOVA design, in which we compared three independent groups on a single factor. Repeated-Measures ANOVA has a corollary design, though it does not have a special name, in which we can compare three (or more) repeated observations taken from the same group under different circumstances or over time. The example we will use to illustrate this design is one in which the researcher is interested in comparing student satisfaction ratings over time, as they progress from freshman to sophomore, junior and

finally, senior status. Note that in this design, I am measuring student satisfaction among the same group of students, once per year for four years. By definition, this is a hypothetical longitudinal study that requires all subjects to proceed through the repeated observational levels before the final analysis can begin.

In addition to the ANOVA assumptions discussed in the beginning of this chapter, note that for RM-ANOVA designs, subjects must provide data for each repeated observation level in order to be included in the analysis. This can represent a disadvantage of utilizing such an approach, as subject attrition can become a serious issue in longitudinal studies. Nevertheless, assume that we have a sample of n=30 students who provided ratings of satisfaction with some aspect of student learning or living at their university on four separate occasions (as freshmen, sophomores, juniors, and seniors). Figure 8 shows the SPSS output of a RM-ANOVA analysis of these hypothetical student evaluations. Because this design has no independent factors, Levine's test for homogeneity is not appropriate. However, RM-ANOVA has a similar assumption that the variance across pairs of observation be constant. Mauchly's test of sphericity tests the null-hypothesis of constant pairwise variance, and here we have a non-signficant Mauchly's, indicating that our data meet this assumption.

Also notice that the output for the RM-ANOVA statistic produces two ANOVA summary tables—one for repeated-measures factors (called "Within-Subjects" on SPSS output), another for independent-measures factors (called "Between-Subjects" on SPSS output). Our design only has a single repeated-measures factor, so we can ignore the between-subjects ANOVA table for now. Because we met the sphericity assumption in this data, we can interpret the rows labeled "sphericity assumed" from the within-subjects ANOVA summary table. Thus, the overall effect for students' satisfaction ratings changing over time is shown by the $F = 23.7$, $p = .000$, a significant result. This provides statistical evidence that student perceptions are changing significantly over time; however, it does not provide the details about pair-wise comparisons of student satisfaction.

Caveate: Using Apriori Contrasts with Repeated-Measures Designs

In our discussion of the IM-ANOVA, we addressed using Post-Hoc comparisons (such as the Tukey's HSD, Games-Howell, LSD, and others) as a logical extension to a significant omnibus F-test, in order to determine which pairs of groups differed statistically from one another. Repeated-Measures designs cannot take advantage of the commonly recognized post-hoc statistical tests because they fail to account for the within-subject correlation across multiply repeated observations. However, there are statistical alternatives to the post-hoc test for repeated measures designs, called "apriori contrasts."

Like post-hoc tests, apriori contrasts are used to make multiple

Figure 8
Single-Factor Repeated-Measures ANOVA on Student Satisfaction during Four Repeated Observations

Descriptive Statistics

	Mean	Std. Deviation	N
FRESH	6.77	1.569	30
SOPH	4.67	1.561	30
JUNIOR	4.13	1.137	30
SENIOR	6.47	1.717	30

Mauchly's Test of Sphericity [b]

Measure: MEASURE_1

Within Subjects Effect	Mauchly's W	Approx.	df	Sig.	Epsilon[a]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
TIME	.769	7.270	5	.202	.842	.929	.333

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b. Design: Intercept
Within Subjects Design: TIME

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
TIME	Sphericity Assumed	153.025	3	51.008	23.703	.000
	Greenhouse-Geisser	153.025	2.527	60.563	23.703	.000
	Huynh-Feldt	153.025	2.788	54.891	23.703	.000
	Lower-bound	153.025	1.000	153.025	23.703	.000
Error(TIME)	Sphericity Assumed	187.225	87	2.152		
	Greenhouse-Geisser	187.225	73.275	2.555		
	Huynh-Feldt	187.225	80.846	2.316		
	Lower-bound	187.225	29.000	6.456		

Tests of Within-Subjects Contrasts

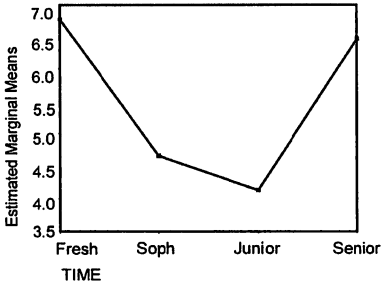
Measure: MEASURE_1

Source		Type III Sum	df	Mean Square	F	Sig.
TIME	Linear	3.082	1	3.082	1.361	.253
	Quadratic	147.408	1	147.408	59.092	.000
	Cubic	2.535	1	2.535	1.494	.231
Error(TIME)	Linear	65.668	29	2.264		
	Quadratic	72.342	29	2.495		
	Cubic	49.215	29	1.697		

Tests of Between-Subjects Effects

Measure: MEASURE_1
Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	3641.008	1	3641.008	1358.207	.000
Error	77.742	29	2.681		



comparisons across the levels of a repeated-measures observation, while at the same time holding experiment-wise alpha risk = .05. There are different types of contrasts, and they test different apriori hypotheses about patterns of change over time. SPSS and other software manufactures make the most commonly used contrasts available from their "point and click" menu systems, and because of page limitation constraints, I will limit my discussion of contrasts to these common choices.

However, it is possible to create custom contrasts that address other hypotheses, and the reader should review the latest syntax guides for their preferred statistical software if custom contrasts are necessary. Note that apriori contrasts can also be used to test for differences across multiple levels of an *independent*-measures factor, but there are limitations to these procedures that make them less popular choices for IM-ANOVA applications. Most notably, apriori contrasts do not make all possible pairwise comparisons across the levels of a factor, but instead test for pre-determined difference, or predicted patterns of differences among levels of the factor. Many times the institutional researcher would rather choose to compare all possible pairs of a factor, and so in IM-ANOVA situations, often choose post-hoc pairwise tests over apriori contrasts.

Because our current example is a RM-ANOVA design, we must rely on apriori contrasts as a follow-up to a significant omnibus F-test. In our current example, I have chosen polynomial contrasts because I am most interested in learning about the trend of change over time, and did not have specific ideas or hypotheses about how satisfaction might change. Polynomial contrasts are appropriate in instances where one wants to capture the overall pattern of change in predictable ways. These contrasts test for the fit of polynomial functions to the repeated observations in increasing complexity, starting with a simple linear function, then testing the quadratic, cubic, quartic, and potentially more complex geometric polynomial functions depending on the number of repeated observations included in the model. Polynomial contrasts test for k-1 number of contrasts, beginning with linear, quadratic, cubic, and so on—each increasing in polynomial complexity. For example, three polynomial contrasts can be tested on a study with k=4 repeated observations (linear, quadratic, cubic). Four polynomial contrasts can be tested on a study with k=5 repeated observations (linear, quadratic, cubic, quartic).

As Figure 9 shows, our linear and cubic trend contrasts are not significant, but the quadratic trend contrast is ($F = 59.05$, $p = .000$). The means chart at the bottom of Figure 7 illustrates this effect nicely, where student satisfaction starts high, dips during sophomore and junior years, then rises again as seniors—the typical "U-shaped" quadratic function. With this information, we not only know that satisfaction changes over time (evidenced by our Omnibus F-test), but also we have an understanding of just *how* student satisfaction changes during their time at the university.

There are other commonly used apriori contrasts that we could have chosen, depending on our original research goals. If we were interested in comparing student performance in a "typical" classroom environment, for example, to the same sample of students whose performance is also assessed in two or more "experimental" classroom environments, we would choose apriori contrasts that compare all experimental environments to the typical classroom using *simple* contrasts, with the typical condition serving as the reference condition. Or, if we had an apriori hypothesis that required a comparison of environment to the mean of the other environments, we would choose *Helmert* or *reverse Helmert* contrasts. Finally, if we hypothesized statistically significant difference on adjacent levels of a repeated measure, the *repeated* contrasts can test our hypothesis. If the researcher wanted to test an apriori hypothesis that doesn't lend itself to these commonly used methods, modern statistical software offers the ability to set up custom contrasts that can make the necessary comparisons while holding the alpha risk to .05, though there are restrictions concerning the number of comparisons the researcher can make, which depends on the number of levels in the repeated measures factor. Custom contrasts are not particularly difficult to create using SPSS, SAS and other software packages; however, it usually requires using the syntax method of running statistical analyses instead of the more popular menu-driven graphic user interface. Generally the user is restricted to a number of comparisons equal to the number of levels/groups-1, and as mentioned previously, this restriction is a major difference between apriori and post-hoc comparisons.

Using Single-Factor Repeated-Measures ANOVA Where Condition is the RM-Factor

In the previous RM-ANOVA example, students were measured repeatedly several times to determine whether self-reported satisfaction changes over time. In that example, *time* was the repeated measures factor. Repeated-Measures ANOVA can also be used to test hypotheses comparing some continuously measured outcome across different *conditions*. Note that we are still discussing experimental designs in which the *same* subjects are measured repeatedly. However in this next example, the repeated measurements are taken under different conditions so that we can establish whether or not there is statistical evidence that these conditions produce different results. This type of design is most similar to the Independent-Measures experimental designs that compare outcomes generated from different groups of subjects representing the different conditions, with the important distinction that with RM-designs, the *same* subjects are assessed.

This next example uses data collected to assess a new teaching/learning paradigm for teaching students how to interpret information embedded on a microscopic biological sample. These data were collected from students who were undergoing laboratory instruction on how to interpret such

information. Students historically learn how to read biological samples by instruction and practice using traditional microscope labs; however, there are also novel computer software programs that effectively mimic what a student would see through a traditional microscope on a computer terminal. We wanted to compare several aspects of student learning and perceptions through the use of the computer-simulated microscope to their learning and perceptions using a real microscope. For several reasons, we believed that students would enjoy using the computer more than the microscope, leading to more practice efforts on the computer, and ultimately better learning. To test our hypotheses, n=114 students were exposed to both methods of instruction for half of a quarter, after which we measured several aspects of student preference and student learning. We compared student outcomes using several Repeated-Measures ANOVA statistic, one per outcome measure.

We believed that a positive learning experience would ultimately lead to better student learning, so we asked students about their laboratory experiences at the end of each half-quarter course. Student responses to several Likert-type questions assessing how much they liked using the traditional microscope and the computer simulations were averaged to create our outcomes measures for this analysis. The results of this analysis are presented in Figure 9, where the data show significantly more positive student experience using the computer-simulated microscope relative to the traditional microscope ($F = 4.6$, $p = .034$). We compared other aspects of student enjoyment and student learning, and consistently found more positive outcomes associated with the computer-simulated microscope experience relative to the traditional microscope. Our analyses of these outcomes were similar to the above analysis, that student perceptions and learning outcomes consistently favored the simulated learning modality over the traditional microscope. Such consistent results on different measures of student perceptions and learning can (and did) make for a compelling case for curriculum change.

Note that this design only compared two conditions, thus we could have used a repeated-measures t-test for this analysis. In reality, our statistical design was a bit more complicated than illustrated here; however, the data set lends itself nicely into a comparison of a repeated-measures t-test to RM-ANOVA. Note that the bottom half of Figure 7 also shows the results of repeated-measures t-test on these data. In situations like this, when there are only two levels of the factor of interest, the F-statistic will be equal to the square of the t-statistic, and the probability values associated with each statistic will be the same. I present this merely as an illustration that either the statistic is appropriate, and although the statistical calculations are different, the statistical significance is the same, and thus the same hypothesis testing decision will be made.

Figure 9
Results of Two Statistical Techniques for Comparing Student Experiences Under Two Different Learning Modalities: Repeated Measures ANOVA versus Repeated Measures T-Test

Descriptive Statistics

	Mean	Std. Deviation	N
My experience w/scope was positive	5.43	1.58	114
My experience with computer was positive	5.67	1.46	114

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
LEARNBY	Sphericity Assumed	3.197	1	3.197	4.614	.034
	Greenhouse-Geisser	3.197	1.000	3.197	4.614	.034
	Huynh-Feldt	3.197	1.000	3.197	4.614	.034
	Lower-bound	3.197	1.000	3.197	4.614	.034
Error(LEARNBY)	Sphericity Assumed	78.303	113	.693		
	Greenhouse-Geisser	78.303	113.000	.693		
	Huynh-Feldt	78.303	113.000	.693		
	Lower-bound	78.303	113.000	.693		

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	My experience w/scope was positive - My experience with computer was positive	-.24	1.177	.110	-.46	-.02	-2.148	113	.034

Factorial Repeated-Measures ANOVA Designs

As with Independent-Measures ANOVA, Repeated-Measures ANOVA can easily be expanded to address experimental designs with two, three, or more repeated-measures factors, though such designs are less common. And similar to IM-ANOVA, these factorial designs provide an evaluation of the main effects for each repeated-measures factor, and the interaction effects involving them. These designs are sometimes referred to as nested designs, higher-order designs, or doubly repeated-measures designs, because the repeated observations are hierarchical in nature.

The approach to higher-order RM-ANOVA analyses is similar to our earlier discussion of higher order IM-ANOVA designs. We concern ourselves first with any significant interaction effects that emerge from the model, running follow-up analyses where justified, until we have a thorough understanding of the multifactor effects on the outcome.

Multifactorial designs involving only repeated-measures factors are less common than completely independent-measures designs, or mixed-model

designs that involve independent- and repeated-measures factors, yet the strategy for analyzing the effects in these designs remains essentially the same. I will move next to a mixed-model ANOVA example.

Mixed-Model ANOVA Designs

The term mixed-model refers to experimental designs that involve at least one independent-measures factor, and at least one repeated-measures factor. These designs are common in the educational and social sciences literature and the strategy for analyzing these hybrid designs is consistent with our discussions of two-factor and higher-order designs discussed above. To illustrate, consider a two-factor mixed-model design involving the independent-measure factor comparing alumni who graduated from one of two majors (Pre-Medicine, Engineering), crossed with a repeated-measure factor of multiple self-reported assessments reporting intellectual growth over time since graduation (three, five, seven years). Thus we have a 2 (major: Pre-Med, Engineering) x 3 (Time: 3,5,7 years since graduation) mixed-model ANOVA design with one independent-measures factor (major) and one repeated-measures factor (time since graduation). The outcome measure in this hypothetical study is alumni self-reported measures of intellectual growth. Assume that we collected these repeated self-assessments from n=25 Pre-Med alumni, and n=25 Engineering alumni.

The results of this two-factor, mixed-model design are shown in Figure 11[2]. Note that because we have both repeated- and independent-measures factors, both of the ANOVA summary tables contain relevant information for our consideration. However, any interactions involving a repeated-measures factor will be shown in the within-subjects ANOVA summary table, so it is there that we will focus our attention first.

In this example, note that we have a significant Years (since graduation) * Major interaction effect ($F = 4.92$, $p = .009$). We also have a significant effect for Years since graduation, and a non-significant Major effect but these effects are qualified by the significant interaction term in our model, suggesting the need to further reduce our model.

The significant interaction term tells us that the pattern of changing self-reports of intellectual growth over time for Pre-Med graduates is not the same as the pattern of changing intellectual growth for Engineering graduates (see line chart in Figure 10a). As discussed previously, one school of statistical thought would be to stop here and interpret the means data accordingly. Alternatively, with a significant interaction effect justifying a subsequent analysis, we could hold one factor constant and examine the simple-effects on the remaining factor. In this case, I chose to examine the simple-effects of time on intellectual growth, holding major constant (Figure 10b). The simple-effects analyses revealed no significant differences on intellectual growth between three, five and seven years since graduation for Engineering alumni. However, Pre-Medicine graduates' self-reported significantly higher

Figure 10a
Two Factor Mixed-Model ANOVA Comparing Self-Reported Intellectual Growth by Major and Over Time Since Graduation

Descriptive Statistics

	MAJOR	Mean	Std. Deviation	N
Intel Growth @ 3 yrs.	Pre-Med	6.0851	1.13133	25
	Engineering	7.2071	1.78875	25
	Total	6.6461	1.58593	50
Intel Growth @ 5 yrs.	Pre-Med	6.8926	1.60953	25
	Engineering	6.7837	1.90866	25
	Total	6.8381	1.74820	50
Intel Growth @ 7 yrs.	Pre-Med	8.3640	1.07089	25
	Engineering	7.4968	1.54049	25
	Total	7.9304	1.38416	50

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
YEARS	Sphericity Assumed	47.990	2	23.995	9.368	.000
	Greenhouse-Geisser	47.990	1.993	24.081	9.368	.000
	Huynh-Feldt	47.990	2.000	23.995	9.368	.000
	Lower-bound	47.990	1.000	47.990	9.368	.004
YEARS * MAJOR	Sphericity Assumed	25.197	2	12.598	4.919	.009
	Greenhouse-Geisser	25.197	1.993	12.643	4.919	.009
	Huynh-Feldt	25.197	2.000	12.598	4.919	.009
	Lower-bound	25.197	1.000	25.197	4.919	.031
Error(YEARS)	Sphericity Assumed	245.895	96	2.561		
	Greenhouse-Geisser	245.895	95.660	2.571		
	Huynh-Feldt	245.895	96.000	2.561		
	Lower-bound	245.895	48.000	5.123		

Tests of Between-Subjects Effects

Measure: MEASURE_1

Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	7643.104	1	7643.104	3833.694	.000
MAJOR	.089	1	.089	.044	.834
Error	95.696	48	1.994		

Estimated Marginal Means of MEASURE_1

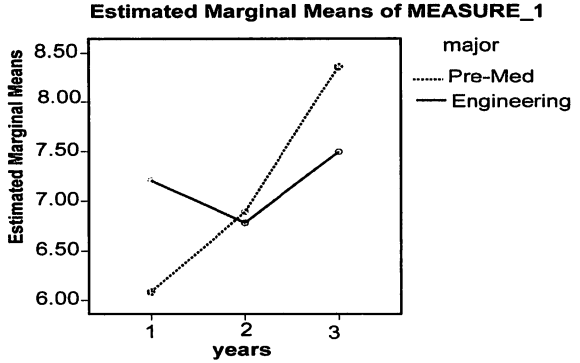


Figure 10b
Follow-Up Simple-Effects Repeated-Measures ANOVA Results on Intellectual Growth by Time Within Level of Major

Descriptive Statistics

	MAJOR	Mean	Std. Deviation	N
Intel Growth @ 3 yrs.	Pre-Med	6.0851	1.13133	25
	Engineering	7.2071	1.78875	25
	Total	6.6461	1.58593	50
Intel Growth @ 5 yrs.	Pre-Med	6.8926	1.60953	25
	Engineering	6.7837	1.90866	25
	Total	6.8381	1.74820	50
Intel Growth @ 7 yrs.	Pre-Med	8.3640	1.07089	25
	Engineering	7.4968	1.54049	25
	Total	7.9304	1.38416	50

Tests of Within-Subjects Effects

Measure: MEASURE_1

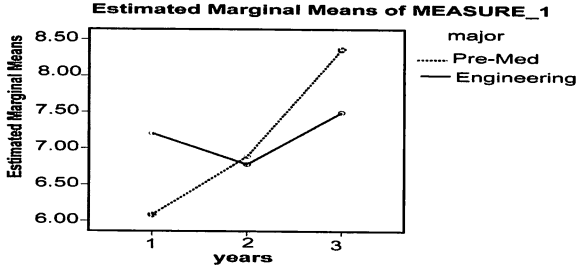
Source		Type III Sum of Squares	df	Mean Square	F	Sig.
YEARS	Sphericity Assumed	47.990	2	23.995	9.368	.000
	Greenhouse-Geisser	47.990	1.993	24.081	9.368	.000
	Huynh-Feldt	47.990	2.000	23.995	9.368	.000
	Lower-bound	47.990	1.000	47.990	9.368	.004
YEARS * MAJOR	Sphericity Assumed	25.197	2	12.598	4.919	.009
	Greenhouse-Geisser	25.197	1.993	12.643	4.919	.009
	Huynh-Feldt	25.197	2.000	12.598	4.919	.009
	Lower-bound	25.197	1.000	25.197	4.919	.031
Error(YEARS)	Sphericity Assumed	245.895	96	2.561		
	Greenhouse-Geisser	245.895	95.660	2.571		
	Huynh-Feldt	245.895	98.000	2.561		
	Lower-bound	245.895	48.000	5.123		

Tests of Between-Subjects Effects

Measure: MEASURE_1
Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	7643.104	1	7643.104	3833.694	.000
MAJOR	.089	1	.089	.044	.834
Error	95.696	48	1.994		

Estimated Marginal Means of MEASURE_1



Note that SPSS is representing the levels of the repeated-measures factor (years) as 1,2 and 3, representing three, five, and seven years post-graduation.

intellectual growth between years three and seven post-graduation ($p = .000$). Judging from the means data, this intellectual growth appears to be fairly linear (see line chart on Figure 10). From these analyses we can conclude that Pre-Medicine graduates report a linear increase in intellectual growth from three to seven years post-graduation, resulting in significantly higher self-reported intellectual growth at year seven compared to year three, yet Engineering graduates report relatively consistent intellectual growth over this time period.

Using Covariates in Factorial ANOVA Designs

By now the reader should recognize the power and flexibility of ANOVA to handle both independent- and repeated-measures factors, in combination or alone, and to address virtually any factorial design for hypothesis testing. We have illustrated through numerous examples how one can apply the ANOVA statistic to answer fairly complex questions about differences between and within groups, over time and across multiple conditions. ANOVA is also capable of incorporating continuous measures into the statistical model in order to adjust for potential differences on this measure in the evaluation of the effects on the other factors in the model. It is important to recognize the difference between factors and covariates. Factors are categorical ways of distinguishing between groups of subjects (e.g., gender, male vs. female) or between distinct observational time periods (e.g., pre vs. post), whereas, a covariate is something that is measured on a continuous scale (e.g., age). In ANOVA, including covariates in a model is typically done to remove the potential effects of the covariate on the outcome, so that a more sterile assessment of the main effects and interactions involving the factors is possible. Incorporating a covariate in an ANOVA design is a means for statistically removing any effects attributable to the covariate so that the "adjusted means" can be evaluated against the factors in the model.

To illustrate an ANOVA using a covariate, recall our earlier example that compared faculty salaries by department, tenure status, and gender. Our simple-effects analysis of the Biology department (Figure 6) showed significant gender differences for both tenured and untenured faculty. However, the Biology chair might be interested in determining whether or not the *difference* between male and female salaries for tenured faculty (those who have been faculty members longer than their untenured cohorts) is as dramatic as the gender difference in untenured faculty. Such an inquiry might provide more insight into the changing trends in the salary inequity, assuming that one accepts the argument that a salary bias in tenured faculty represents biases that initially occurred (with starting salaries) some time ago relative to the salary inequity observed in the more recently hired untenured faculty. In order to explore this possibility, we need to run a two-factor ANOVA (Gender * Tenure Status) on only the Biology faculty members' salaries. If this analysis reveals a significant Gender * Tenure interaction effect, then evidence exists that supports the idea that the gender bias for tenured faculty is not the same as the bias in untenured faculty. Figure 11a shows the results of this analysis, with a significant Tenure * Gender interaction effect ($F = 7.97$, $p = .008$) showing a greater gender-difference in salaries among tenured faculty relative to untenured faculty in this department. Even with this result, our earlier simple-effects analysis confirmed that there remains a statistically significant difference between male and female salaries in both the tenured and untenured faculty, and even if there is "less of a difference" among untenured faculty members, that's not particularly satisfying in terms of gender equity.

Figure 11a
Two-Factor ANOVA on Faculty Salary by Tenure Status and Gender in the Biology Department Only

Descriptive Statistics (Biology Dept. Data Only)

Dependent Variable: salary

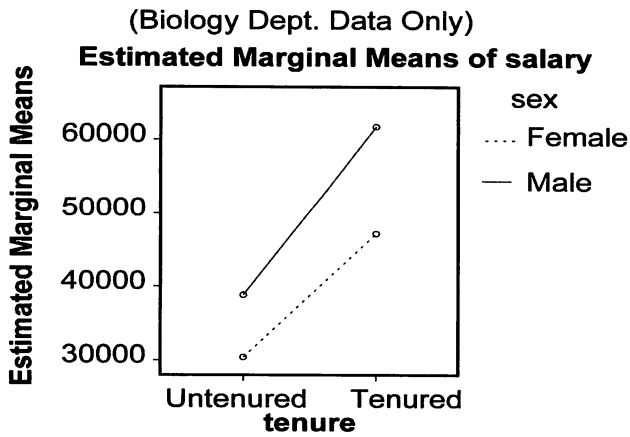
sex	tenure	Mean	Std. Deviation	N
Female	Untenured	30405.30	2924.151	10
	Tenured	47142.60	1392.477	10
	Total	38773.95	8870.688	20
Male	Untenured	38834.20	4568.788	10
	Tenured	61705.30	3979.357	10
	Total	50269.75	12451.629	20
Total	Untenured	34619.75	5712.646	20
	Tenured	54423.95	8014.230	20
	Total	44521.85	12155.452	40

Tests of Between-Subjects Effects

Dependent Variable: salary

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	5337656309[a]	3	1779218770	150.785	.000
Intercept	7.929E+10	1	7.929E+10	6719.482	.000
sex	1321534176	1	1321534176	111.997	.000
tenure	3922063376	1	3922063376	332.387	.000
sex * tenure	94058756.1	1	94058756.10	7.971	.008
Error	424788814	36	11799689.28		
Total	8.505E+10	40			
Corrected Total	5762445123	39			

a. R Squared = .926 (Adjusted R Squared = .920)



One might speculate that the gender bias in this department is possibly due to *age differences* among our faculty, not purely gender, and that it so happens that the higher-paid faculty happen to be older and more experienced, thus deserving of higher salaries. To test this hypothesis, we will run the same 2 (Gender) x 2 (Tenure Status) IM-ANOVA, but we also include faculty AGE into the model as covariate. The results of this IM-ANOVA with covariate are shown in Figure 11b. Note that this model compares *adjusted salaries*, where Age is used as the covariate adjustment factor. Thus it is testing the more sterile effects of tenure status and gender on salary, after the effects of Age

Figure 11b
Two-Factor ANOVA on Faculty Salary by Tenure Status and Gender, using Age as a Covariate

Tests of Between-Subjects Effects

Dependent Variable: SALARY

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	5.440E+09[a]	4	1360098328	147.813	.000
Intercept	222380733	1	222380733.0	24.168	.000
AGE	102737002	1	102737001.7	11.165	.002
SEX	12055026.2	1	12055026.21	1.310	.260
TENURE	361795367	1	361795366.5	39.319	.000
SEX * TENURE	285308.417	1	285308.417	.031	.861
Error	322051813	35	9201480.358		
Total	8.505E+10	40			
Corrected Total	5.762E+09	39			

a. R Squared = .944 (Adjusted R Squared = .938)

1. SEX

Dependent Variable: SALARY

SEX	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Female	42980.03[a]	1429.877	40077.221	45882.829
Male	46063.67[a]	1429.877	43160.871	48966.479

a. Covariates appearing in the model are evaluated at the following values: AGE = 39.10.

2. TENURE

Dependent Variable: SALARY

TENURE	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Untenured	37814.53[a]	1172.268	35434.697	40194.360
Tenured	51229.17[a]	1172.268	48849.340	53609.003

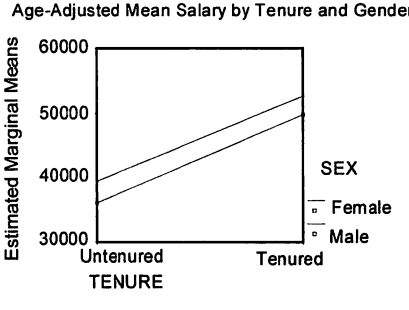
a. Covariates appearing in the model are evaluated at the following values: AGE = 39.10.

3. SEX * TENURE

Dependent Variable: SALARY

SEX	TENURE	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Female	Untenured	36151.30[a]	1969.067	32153.886	40148.724
	Tenured	49808.75[a]	1247.716	47275.748	52341.744
Male	Untenured	39477.75[a]	978.387	37491.521	41463.984
	Tenured	52649.60[a]	2874.868	46813.304	58485.889

a. Covariates appearing in the model are evaluated at the following values: AGE = 39.10.



on salary have been removed. Note from the ANOVA summary table, that Age is significantly related to salary, and now that Age is part of our ANOVA model, we no longer have a significant Gender * Tenure interaction term, or a main effect for Gender. The only significant effect remaining is a main effect for Tenure status, showing significantly higher age-adjusted salaries for tenured versus untenured faculty ($F = 39.3$, $p = .000$). This is a very different conclusion than we arrived at in our original discussion of these salary data and illustrates the importance of including relevant covariates in ANOVA models, particularly when dealing with issues as important as salary equity and discrimination. The next important task would be to ponder why male faculty in this department are older than women, and the bearing this has on gender equity. Perhaps men got an earlier start in this discipline than women, indicating that the gender bias is mediated by age. Or, perhaps a gender bias existed many years ago, when the older faculty members were hired, but that more recent hiring policies have made strides to correct for the bias. We cannot devote too much time to understanding these results (and they are from fictitious data anyway), but the point remains that one must consider relevant covariates in any analytic approach, and ANOVA can incorporate continuous covariates in this manner.

Note that this example illustrates a situation in which a covariate (age) was responsible for effects that could have falsely been attributed to other factors (gender) in the model. Sometimes covariates have no effects whatsoever on the significance of other factors, and it is possible that including a covariate in a model could increase the likelihood of significance on other factors already included in the model. The decision of whether or not to include covariates in an ANOVA model should, therefore, be based on a logical, defendable argument. It is not "good science" to create expansive statistical models, with multiple factors and covariates, just because it is possible to do so. One must carefully and critically evaluate each factor and covariate that is to be included in a statistical model, and only accept those that are deemed important and statistically relevant.

Presenting Results of ANOVA Models

By now, the reader should be well informed enough to understand when the ANOVA statistic might be used to shed light on Institutional Research and Assessment topics of interest. Hopefully one also has a good idea of the different types of ANOVA models that can help the researcher understand their data better, and how incredibly versatile and powerful this statistic is. By now the reader should have an appreciation for the fact that complex, multi-factorial ANOVA models have similarly complex results that are perhaps not so intuitively easy to understand without some guidance and practice. Unfortunately, it is often the case that we as professionals have far more statistical training and practice than our colleagues to whom we must present our results. As such, in addition to having a solid foundation in statistical theory and practice,

it is possibly equally as important that we refine our presentation skills so that we can make complex (and interesting!) results seem simpler to understand. Thank heavens for graphs and presentation software!

There are many resources available that attempt to teach us how to understand, evaluate, and present statistical results appropriately (see Abelson, 1995, van Belle, 2002, and Farebrother, 2002 to name a few). No better tool to ease presentation exists then a solid understanding of the statistical procedure that you ran, and the results obtained. With regard to presenting the results from complex statistical models, like a multi-factorial ANOVA, we have generally found that pictures speak volumes over tables of numbers, and this is especially true for non-quantitative audiences. The reader need only to flip back a few pages and compare their own "knowledge-gained" from looking at the Descriptive Statistics tables that are produced as part of the SPSS output, to how much better they understand the effects when they look at the line charts that accompany many of the figures in this chapter.

It's not that clearly defined tables are not valuable—they are—especially in written reports that an audience may refer to sometime after a visual/audio presentation. But while tables of numbers may speak to some, pictures seem to somehow shout to everyone. So whenever possible, use a variety of different bar charts or line graphs to illustrate where significant differences exist. For significant interaction effects, I highly recommend use of line charts over bar graphs, as it is easy to see when lines are parallel (indicating no interaction effect, see Figure 11b) versus when they "spread apart" from one level of a factor to the next (indicating an interaction effect, See Figure 11a), or even better, when they cross over from one level of a factor to the next (the strongest interaction effect possible).

Gerald van Belle, in his highly acclaimed 2002 textbook titled *Statistical Rules of Thumb* (an intelligent, yet hilarious read for anyone in our field) devoted several pages to the notion of which types of charts best illustrate statistical effects. Dr. van Belle tells us to "always graph the data," but to "never use a pie chart." He directs further that "bargraphs waste ink [and] don't illuminate complex relationships," and that "stacked bargraphs are worse than bargraphs." Finally, in his own humorous style, he writes that "three-dimensional bargraphs constitute misdirected artistry!" The point in all of this is that one should be careful not to get too wrapped up in the fancy charting options that SPSS, PowerPoint or other presentation software offers, but to focus more on the types of charts that show significant *differences* that exist in your data, and/or *interaction effects* that have been revealed. At the same time, be weary of the default scaling that these software products use when graphing data, as the scales are almost never what you want them to be.

Summary Remarks

This chapter began with a discussion of why the Analysis of Variance statistic is a valuable analytic tool, starting off with a simple extension to the Independent Measures t-statistic with more than two comparison groups, and working through increasingly complex multifactorial experimental designs. We presented several examples relevant to Institutional Research and Assessment offices, where ANOVA could be applied in studies utilizing Independent-Measures factors, Repeated-Measures factors, and Mixed-Model designs. We discussed a general analytic strategy for understanding higher-order interaction effects, and how to dig deeper into these effects with follow-up post-hoc and/or apriori contrast analyses. We then shifted our discussion of ANOVA to an illustration of how the statistic can also incorporate covariates into the model in order to better understand the effects of model factors on outcomes, after removing the effects of a covariate.

Before ending our discussion of ANOVA, it is important to again emphasize that the ANOVA statistic is a very versatile and powerful tool for understanding complex effects in multifactorial experimental designs—so much so that I feel it necessary to caution the reader against the temptation of creating what I call the "kitchen sink" study, only to then have to somehow interpret complex higher-order interactions that are very difficult to grasp. A "kitchen sink" study is one that includes too many factors, too many covariates, or both. While our statistical software can and will crank out results for such broadly defined studies, the result is often a significant three, four, five or higher-order interaction effect that is, for all practical sense, impossible to understand. Even if the eager analyst digs deep in his or her understanding of such effects, ultimately the researcher is going to have to *explain* these complex interaction effects to higher education administrators and decision-makers, who often are not as well-equipped, interested or motivated in understanding complex statistical effects resulting from "kitchen sink" studies. Therefore, for statistical, theoretical, and practical reasons, the use of multifactorial ANOVA designs to address Institutional Research and/or Assessment questions should be conducted in a thoughtful manner. Always strive for parsimonious models that can provide concrete answers to simple questions, rather than overly complex models that ultimately confuse and convolute our understanding of what is most likely a much simpler question.

References

Abelson, R. P. (1995). *Statistics as principled argument.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Coughlin, M. A., & Pagano, M. (1997). *Case study applications of statistics in institutional Research. Resources in Institutional Research, Number Ten.* Tallahassee, FL: Association for Institutional Research.

Farebrother, R. W. (2002). *Visualizing statistical models and concepts.* New York: Marcel Dekker.

Rosenthal, R., & Rosnow, R. L. (1991). *Essential of behavioral research: Methods and data analysis (2nd Ed.).* New York: McGraw-Hill Series in Psychology.

Tabachnick, B.G., & Fidell, L. S. (1989). *Using Multivariate Statistics (2nd Ed).* New York: Harper Collins.

van Belle, G. (2002). *Statistical rules of thumb.* Canada: Wiley-Interscience Series in Probability and Statistics.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *statistical principles in experimental design.* New York: McGraw-Hill Series in Psychology.

Endnotes

[1] One can use ANOVA if only two groups exist, though the t-test is generally preferred because it has fewer assumptions, and is easier to explain to a statistically naive audience. In fact, the calculation of an F-statistic comparing two groups would equal the square of a t-score comparing the same two groups, each with exactly the same *p*-value.

[2] I eliminated some output from this Figure because of space limitations. These data passed all assumptions of ANOVA.

Chapter 3
Regression Analysis for Institutional Research

Robert K. Toutkoushian

This chapter will review the statistical technique known as multiple regression analysis, and how it can be used to examine problems commonly faced in institutional research. The previous chapters in this monograph have explored statistical techniques for decomposing the means between different groups of observations. In contrast, the basic objective of regression analysis is to estimate the linear relationship between a set of K independent variables (denoted X_1 to X_K) and a specific dependent variable (denoted Y). In regression analysis, the dependent variable is assumed to be continuous over the range of values under consideration. When the dependent variable is dichotomous (0,1), a related technique known as logistic regression analysis can be used to accomplish the same goal. This topic will be covered in chapter five.

One of the main strengths of regression analysis as a statistical technique is its flexibility. Regression analysis can be used when the independent variables are continuous, discrete, or dichotomous, and thus can handle most of the different types of variables used in institutional research. By appropriately transforming the dependent and/or independent variables in the model, regression analysis can also estimate a number of non-linear as well as linear relationships between variables of interest.

Regression analysis has three main uses that are important for institutional research applications. The first is that the model can be used to test hypotheses regarding the relationships between specific independent variables and a given dependent variable. This is important because most of the situations encountered in institutional research are those where the analyst does not observe the data-generating process. Through hypothesis tests, it is possible to draw inferences as to whether an independent variable of interest (such as a student's grade point average) has an influence on a dependent variable (such as a student's income after graduation) for the entire population.

Regression analysis also permits the analyst to estimate coefficients showing how changes in an independent variable affect the dependent variable. These coefficients can have important policy implications and are important beyond the fact that the independent variable has a significant effect on the dependent variable. For example, the Director of Enrollment Management at an institution may want to know not only if an applicant's SAT score affects his or her future grade point average, but how large the effect is. Finally, regression analysis can be used to derive predictions for the dependent variable under consideration. Once the regression equation has been specified, values of the independent variables can be substituted into the equation and predictions obtained for Y.

There are many situations encountered in institutional research where the use of regression analysis can be extremely valuable. Regression models can be used in a variety of ways to assist a campus in fulfilling its enrollment management functions. For example, a regression model could be used to identify whether factors such as financial aid offers and advertising expenditures affect the number of applications received in a given year. The same model could be used to quantify the change in applications or enrollments that might occur if financial aid offers are increased by specific amounts. This technique could be used to look at how characteristics of applicants, such as family income, student ability, race, gender, and so on, affect the future academic performance of students at an institution. This information could prove to be extremely valuable in making admissions decisions and identifying students who are likely to need remediation assistance.

This chapter deals exclusively on regression analysis and its application to institutional research problems. It is assumed that the reader has some working knowledge of how to conduct a hypothesis test, and understands the notion of correlation and simple (1 variable) regression analysis. The emphasis will be on how to use and adapt the technique to situations that might be encountered in institutional research, rather than derive the mathematical properties of different procedures. The chapter concludes with two examples of how these techniques can be useful for the practice of institutional research.

Basic Regression Model

The basic form of a regression equation can be written as follows:

$$Y_i = \beta_0 + \sum_{j=1}^{K} \beta_j X_{ij} + \varepsilon_i \qquad [3.1]$$

where Y_i = value of the dependent variable for the i-th observation, X_{ij} = value of the j-th independent variable for the i-th observation, and e_i = random error term for the i-th observation. The parameters $\beta_0,\ldots,\beta_K$ are population coefficients to be estimated through the use of ordinary least squares. Each coefficient is referred to as a "partial effect," which means the impact of a particular variable (say X_1) on Y holding constant the other variables in the model. Ordinary least squares specifies that the estimates of these population parameters should be determined by the set that minimizes the sum of squared residuals (errors). Once the coefficients have been estimated (denoted b_0 to b_K), the equation can be written in a linear form. For example, if Y = per-student cost at the i-th college, X_1 = student to faculty ratio at the i-th college and X_2 = total (headcount) enrollment at the college, the resulting regression equation might be written as:

$$\hat{Y} = 4000 - 200 X_1 + 50 X_2$$

where $\hat{Y}$ = predicted value of Y based on the regression equation. In this hypothetical example, $b_0 = 4000$, $b_1 = -200$, and $b_2 = 50$.

Assumptions in Regression Analysis

There are several key assumptions that must be made regarding the multiple regression model:

1. The dependent variable is continuous.
2. The independent variables are uncorrelated with each other.
3. The independent variables are uncorrelated with the error term.
4. The error term has a mean of zero, a constant variance, and the errors are uncorrelated with each other.

It is important to note that regression analysis is fairly robust with regard to small violations of these assumptions. In many empirical studies, the dependent variable may have a large number of possible values and yet not be continuous. Such is the case for income, and the number of credit hours attempted/completed by students. Even concepts which in theory are continuous, such as time, may be measured only in discrete units. Regression analysis can still be used in these situations provided that a continuous distribution is a reasonable proxy for the true distribution.

Turning to the second assumption, in most empirical studies there will be some correlation between the various independent variables in the regression model. Correlation between the independent variables will affect the coefficient estimates because these coefficients are interpreted as partial effects. This can present problems for the institutional researcher who is interested in the effect of a particular factor (such as ability to pay) on student performance in college. If the researcher were to include multiple measures of a student's ability to pay, such as family income and wealth, into the regression equation explaining future academic performance, then the resulting partial effects of family income and wealth on performance would be affected by the likely correlation between them.

On the other hand, there are situations where the analyst will be very interested in capturing this correlation and removing its effects from a particular policy variable. Returning to the previous example, if the analyst wanted to know how student ability affects performance, it would be very important to remove the effects of other factors that would also affect performance and may be correlated with ability, such as family income, parental education, etc. In this instance, it is less problematic that multiple measures of socioeconomic status are used in the regression model and possibly correlate with each other. The main point to take away from this discussion is that

correlation between the independent variables is not always a problem in practice.

A special problem known as multicollinearity can arise, and has serious complications for regression analysis. If two independent variables are collinear, then the ordinary least squares technique will not be able to properly identify the partial effects and standard errors of the two variables. The result is that the coefficients will be biased and the standard errors will be greatly inflated, making it appear as though neither variable has an influence on the dependent variable. While analysts often express concern about the possibility of multicollinearity, it is important to note that this only arises when the two variables have an extremely high correlation with each other. If the two variables are correlated, but not to the point that they are collinear, then ordinary least squares will still be able to calculate correct coefficient estimates and standard errors.

Turning to the third assumption, it is possible that one or more independent variables in the regression model could be correlated with the error term since the error term captures the net effect of all omitted factors from the model. In most applications, the institutional researcher will not have data on all of the relevant characteristics that theory would suggest should influence the dependent variable. In these instances, the partial effect observed in the regression equation may be biased because it does not remove the effects of correlation with the omitted variables. Of all the problems facing analysts when using multiple regression analysis, this is perhaps the most difficult to overcome because the true (population) model is rarely observed and virtually all regression models are subject to criticism for omitting relevant factors.

This problem can also occur when one of the independent variables is itself an endogenous variable. There are potentially many instances in education research where this may occur. For example, a researcher may wish to estimate the influence of a student taking an Advanced Placement (AP) course on future academic performance. It would not be surprising to find that if this variable were regressed on student performance, the analyst would find that students who have taken AP courses perform better than other students in college. However, students are not randomly assigned to AP courses but rather have to choose to enroll in such courses. This decision is likely to be influenced by factors such as the student's academic ability, parental education, and other factors. In other words, the variable is itself endogenous, and failure to take this into account when conducting the analysis could give rise to misleading results.

Finally, the last assumption needed to derive a multiple regression equation is that the mean of the error term is zero and that the errors have a constant variance and are unrelated to each other. By construction, the error term will have a mean of zero provided that an intercept term is added to the equation. The assumption that the variance of the error term is constant can

be violated in instances where the variance is correlated with one or more variables in the model. When this occurs, it is referred to as heteroscedasticity. This problem most often occurs in cross-sectional data (i.e., data collected from observations at one period of time), and can be corrected through several procedures including weighted least squares. The assumption of uncorrelated errors can be violated when using time series data (i.e., data collected on the same observation at multiple periods of time) and is referred to as autocorrelation. Both of these problems – heteroscedasticity and autocorrelation – can affect the estimated standard errors for the coefficients in the model and the hypothesis tests conducted on them. As with heteroscedasticity, there are statistical procedures that can be used to detect and correct for autocorrelation.

Before proceeding, the reader should also take note of what does *not* have to be assumed when performing regression analysis. First, it is not necessary to assume that any of the variables in the model have a particular distribution, such as the normal distribution. Regression analysis can handle situations where the independent variables are continuous, discrete, and even dichotomous. The only requirement is that the dependent variable must be reasonably approximated by a continuous distribution. A second point to note is that it is not necessary to assume that the error term is normally distributed for the purpose of obtaining estimated coefficients. This assumption is needed, however, if the analyst attempts to perform a hypothesis test on any of the coefficients in the model and the sample size is small. As the sample size increases, the set of possible sample coefficients will approach a normal distribution because of the Central Limit Theorem, and thus the assumption of a normally distributed error term is not required for performing hypothesis tests in large samples.

Uses of Regression Analysis

There are a number of ways in which the estimated multiple regression equation can be used to help understand problems faced in institutional research. The first of these is to test hypotheses concerning the effects of particular independent variables on the dependent variable. The second is to assess the overall quality of the regression model and how well it explains the variation in the dependent variable. The third is to use the equation to derive predictions for the dependent variable using assumptions about the values of the independent variables.

Hypothesis Testing

Most readers should be familiar with the basic notion of how to conduct a hypothesis test. The three main steps used in virtually all hypothesis tests are the following:

1. Specify the null (H_0) and alternative (H_A) hypotheses.

2. Identify the appropriate test statistic and find the critical values for this statistic when the null hypothesis is true.
3. Calculate the test statistic and compare it to the critical values.

The main difference across statistical procedures is in the appropriate test statistic and the random variable being examined. In general, the z- or t-test statistic for any random variable takes the following form:

$$\text{Calculated z- or t-statistic} = \frac{\text{Value of random variable} - \text{mean (rv)}}{\text{Standard error (rv)}} \qquad [3.2]$$

Different situations call for the use of different random variables, and each random variable has its own mean and standard error when the null hypothesis is correct.

In the case of regression analysis, the random variable typically under examination is the estimated coefficient for one of the independent variables. When the error term is normally distributed, and/or the sample size is large (30 or more degrees of freedom is a commonly-used rule of thumb), the estimated coefficient b_j will also be normally distributed with a mean of β_j and a standard error that is a function of the sample size, variance of X_j and correlation with other X's in the model. Hypothesis tests in multiple regression are usually described as "significance tests" because the null hypothesis is that the j-th variable has no effect on the dependent variable (H_0: $\beta_j = 0$; H_A: $\beta_j \neq 0$). Accordingly, the t-test for determining whether the j-th variable has a significant effect on the dependent variable is written as:

$$\text{Calculated t-ratio} = b_j / \text{st.err}(b_j) \qquad [3.3]$$

where b_j = estimate of β_j, and st.err(b_j) = standard error of b_j. The calculated t-ratio has N-K-1 degrees of freedom (N=sample size, K=number of independent variables in the model) and follows the Student t-distribution. When the calculated t-ratio exceeds the critical t-ratio, the null hypothesis can be rejected and the analyst can conclude that the j-th variable has an effect on the dependent variable. In large samples, the normal distribution can be used as an approximation for the t-distribution.

Technically, the coefficients in a regression model can be estimated provided that the sample size is greater than the number of parameters to be estimated. However, a word of caution is advised when estimating models with relatively few degrees of freedom. When there are few degrees of freedom in the model, the resulting standard errors of the coefficients will be large, making it more difficult to reject the null hypothesis and conclude that an independent variable has an effect on the dependent variable. Whenever possible, analysts are encouraged to use a data set that is large enough to

provide more reliable estimates of the coefficients in the model. This is not always possible in institutional research applications, especially when dealing with time-series data. In these instances, analysts should use caution when drawing inferences based on their results.

Goodness-of-Fit Measures

The term "goodness-of-fit" describes how well the multiple regression model explains variations in the dependent variable. There are two primary measures of goodness-of-fit that are used by analysts. The first is the F-test which is used to determine if the variables in the regression model collectively explain a significant proportion of variation in the dependent variable. In practice, this is not a very powerful test because most regression models are capable of passing this test, even when the explanatory power of the model is relatively low. Therefore, it is not often used as a strict indicator as to whether or not the overall regression model is "good."

The most commonly used measure of goodness-of-fit is the coefficient of determination, or R^2. The coefficient of determination measures the proportion of deviation in the dependent variable that is explained by deviations in the independent variables in the model. This is computed by dividing the sum of squares explained by the regression equation by the total sum of squared deviations. The value of R^2 must fall between 0 and 1, with $R^2 = 0$ meaning that the equation explains no portion of the deviations in the dependent variable and $R^2 = 1$ indicates that all of the deviations in the dependent variable are accounted for by the regression model. As R^2 increases, the regression model is said to explain a greater proportion of variations in the dependent variable.

Caution should be used when examining the coefficient of determination to evaluate the quality of the overall regression model. First, the value of R^2 will almost always increase as new variables are added to the model because these new variables may capture some additional variation even if the variable itself does not have a significant impact on the dependent variable. While the adjusted R^2 statistic provides information about whether the additional variables lead to an improvement in the model's fit, many analysts prefer the standard R^2 because the adjusted R^2 is not a precise measure of the percentage of deviations in Y explained by the independent variables. A second concern for analysts is that there is no single cutoff point that can be used to determine if the value of R^2 is good. The variation in some dependent variables is much more difficult than others to explain through a regression model. This means that $R^2 = 0.20$ may be relatively low if the analyst was attempting to explain variations in faculty salaries, but might be relatively high if she were examining the teaching evaluation scores of faculty. The best advice for analysts to use when considering the R^2 value in their study is to compare this value to those obtained in studies of similar variables.

Deriving Predictions

The final use of regression analysis is that institutional researchers can use it to obtain predictions of the dependent variable. To derive predictions, the analyst has to insert values for each of the independent variables into the model and then calculate the resulting predicted value of the dependent variable (Y). For example, suppose that an institutional researcher working in an admissions office estimated the following equation relating a student's SAT score (X_1), number of AP classes completed (X_2), and freshmen year grade point average (Y): $Y = 1.20 + 0.001\ X_1 + 0.50\ X_2$. The analyst could then predict that an applicant with an SAT score of 1,200 and three completed AP courses would have a predicted freshmen year grade point average of $Y = 1.20 + 0.001*(1200) + 0.50*(3) = 3.90$. This is referred to as a point estimate. The regression equation can also be used to derive interval estimates for predicted value of the dependent variable that take into account the variability inherent in this prediction. This is particularly useful for conveying to the reader the uncertainty or error that typically accompanies the prediction. Statistical programs such as SPSS will easily generate mean prediction intervals upon request.

Common Variable Transformations

There are many instances in institutional research where one or more of the variables of interest to a researcher cannot be used in their current form. While multiple regression analysis is a very flexible technique, it does require that the variables used in the model be quantitative or non-categorical. Institutional researchers will encounter many problems where the data are alphanumeric or categorical (such as a student's state of residence, gender, or race/ethnicity), yet would like to assess their impact on a particular dependent variable. Likewise, the data may represent responses from survey questions where respondents are asked to rate specific items. While the ratings may be coded as numerical, the underlying variables are in fact categorical (1=strongly agree, 2=agree, etc.). There are also instances where the analyst has good reason to suspect that the nature of the relationship between a particular dependent and independent variable is in fact non-linear, and using the variables in their current form in the regression model would not capture these non-linearities.

In all of these situations, it is possible through variable transformations to use the variables in hand and yet use multiple regression analysis for the study. In the case of categorical data, dichotomous (0,1) variables can be created that would still capture the effects of these variables on the dependent variable. Through appropriate transformations of either the dependent or independent variables, multiple regression analysis can estimate the linear relationship between non-linear variables. The flexibility of multiple regression analysis to handle all of these situations is one of the main strengths of the technique.

Dichotomous Variables

A dichotomous variable (commonly referred to as a "dummy variable") is a numerical variable that has only two possible values: 0 and 1. Dummy variables can be created from most any type of variable by using an assignment rule. The assignment rule describes how the set of values of a variable should be recoded, i.e., which values are assigned the value zero and which values are assigned the value one. In the case of gender, for example, a dummy variable *F* may be created by assigning the value "female" to 1 and the value "male" to 0. This variable could then be included in the regression equation, and the coefficient for the variable would represent the predicted difference in the dependent variable between the two groups. For the purpose of regression analysis, it does not matter whether female = 1 or female = 0 since this will only change the sign and not the significance level or the magnitude of the estimated coefficient.

More than one dummy variable can be created from the categorical variable at the disposal of the analyst. For example, if the analyst had data on the state of residency for students, he or she could create a separate dummy variable for each state (e.g., $S_1 = 1$ if student is from California, 0 otherwise; $S_2 = 1$ if student is from Minnesota, 0 otherwise). The assignment rule can also rely on groups of observations, such as $R_1 = 1$ if student is from a state on the East Coast, 0 otherwise. When creating dummy variables, the number of dummy variables used in a multiple regression model to represent a particular categorical variable must omit at least one category. In the previous example, if the analyst created fifty dummy variables based on the student's state of residence, at least one of these variables would have to be omitted from the regression model. Failure to do so would lead to perfect collinearity between the set of dummy variables and the intercept in the regression equation. It does not matter which variable is omitted from the model, except to note that the coefficients on each variable represent the difference between the category in question and the omitted category on the dependent variable.

Non-Linear Variable Transformations

Although regression analysis is thought of as a linear estimation technique, there are some simple ways using variable transformations to estimate non-linear relationships between variables. This technique greatly increases the flexibility of regression analysis for dealing with problems in institutional research; however, it does increase the difficulty of interpreting results from the regression model. This chapter will describe only two of the most commonly-used transformations useful for applications in institutional research.

Logarithmic Transformations

There are many instances where transforming one or more variables using the natural logarithmic function can be useful for modeling problems

faced in institutional research. The natural log transformation effectively reduces the scale of the variable and is particularly useful in instances where the distribution of the variable is highly skewed to one side or the other. Table 1 illustrates how the natural log transformation changes the values of a given variable:

Table 1
Example of the Natural Log Transformation of a Variable

Original Variable (Y)	Natural Log Transformation (lnY)
1	0
2	0.69
3	1.10
10	2.30
50	3.91
100	4.61
1,000	6.91
10,000	9.21
100,000	11.51

The values in Table 1 show that there is much less variability between the extreme values for the variable created with the natural log transformation than is true for the original variable.

The natural log transformation can be applied to either the dependent variable, independent variable, or both. When it is only applied to the dependent variable, the resulting multiple regression equation is referred to as a semilogarithmic equation:

$$\ln Y_i = \beta_0 + \sum_{j=1}^{K} \beta_j X_{ij} + \varepsilon_i \qquad [3.4]$$

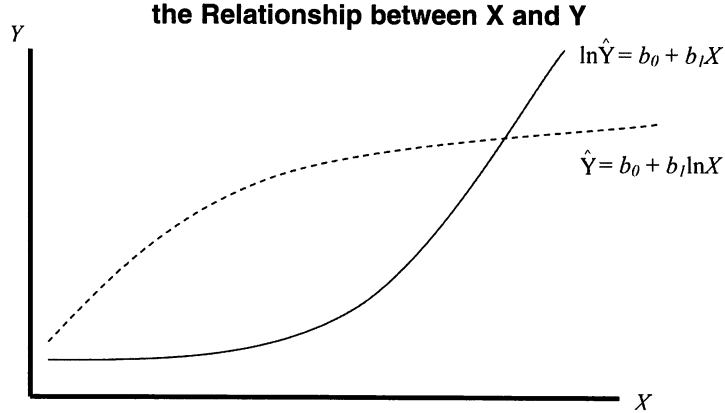
Interpreting this model literally, the estimated coefficients now represent the effect of a one-unit change in each X on the natural log of Y. The coefficients, however, have been shown to be approximations of the percentage changes in Y due to one-unit changes in each X. Therefore, in instances where the analyst has reason to believe that changing the independent variables would have a constant effect on the percentage and not the level of Y, applying the natural log transformation to the dependent variable is very appealing. The transformation implies that Y increases at an increasing rate as X increases, or vice-versa if the relationship between the variables is negative. This transformation is often used in studies of faculty salaries since many institutions of higher education award salary increases on a percentage and not a fixed dollar basis.

Suppose instead that the natural log transformation is made to one of the independent variables in the model, such as:

$$Y_i = \beta_0 + \sum_{j=1}^{K} \beta_j X_{ij} + \beta_{K+1} \ln X_{K+1} + \varepsilon_i \qquad [3.5]$$

The estimated coefficient for β_{K+1} represents the change in Y becuase of a one percent change in X_{K+1}. This transformation implies that Y increases at a decreasing rate as X increases, or vice-versa if the relationship between the variables is negative. This situation may be encountered in institutional research when studying dependent variables that are bounded, such as a student's grade point average or the retention/graduation rate for an institution. Changes in independent variables are likely to have smaller effects on the dependent variable as the upper boundary is approached, and thus the natural log transformation may provide a better fit to the model than would a linear function. Figure 1 shows how these two examples would be represented graphically:

Figure 1
Effect of Natural Logarithmic Transformation on the Relationship between X and Y



Finally, there are times when the analyst may wish to apply the natural log transformation to variables on both sides of the equation. In the resulting equation, the coefficient on X represents the percentage change in Y due to a one percent change in X. This is referred to by economists as an elasticity.

Quadratic

With the natural log transformations described above, the sign of the effect of X on Y, is the same regardless of the value of X. There are times when this is not an accurate description of the relationship between two variables. For example, a student's likelihood of applying to a given college may at first increase as her SAT score increases, but eventually would

decrease as her ability level exceeds the profile of students at the institution. Likewise, studies that have tried to explain per-student costs as a function of enrollments usually posit that per-student costs initially fall as enrollments rise due to economies of scale, but eventually start to increase as the institution becomes too large. In each example, not only does the effect of *X* on *Y* depend on the level of *X*, but the direction of the effect of *X* on *Y* will also change.

To capture these effects in a multiple regression analysis, the analyst creates a new variable that is simply the square of the variable in question and adds this variable to the regression model:

$$Y_i = \beta_0 + \sum_{j=1}^{K} \beta_j X_{ij} + \beta_{K+1} X_{K+1} + \beta_{K+2} X_{K+1}^{\ 2} + \varepsilon_i \qquad [3.6]$$

The effect of X_{K+1} on *Y* can be found by differentiating the resulting equation with regard to X_{K+1}. Likewise, the analyst can determine if there is evidence of a quadratic relationship by applying the standard significance test to the estimated coefficient for the quadratic variable (β_{K+2}). If the estimated coefficient is not statistically different from zero, then there is no evidence of a quadratic relationship between the variable in question and the dependent variable. When $b_{K+2} < 0$, then the quadratic curve increases and then decreases as X increases ("hill-shaped"). Likewise, the curve will be U-shaped when $b_{K+2} > 0$. These two possibilities are shown in Figure 2:

Figure 2
Effect of Quadratic Transformation on the Relationship between X and Y

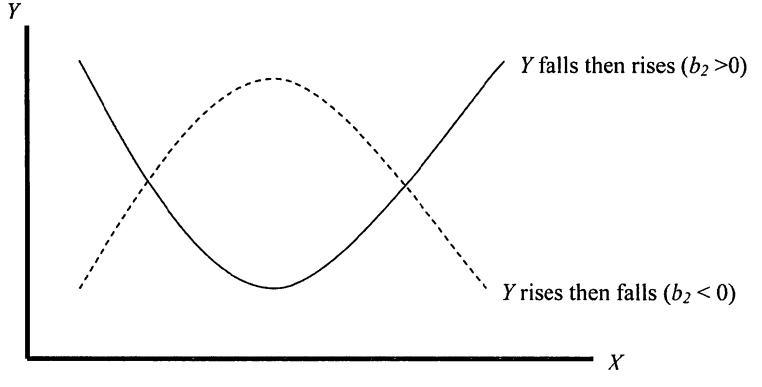


Table 2
SPSS Output: Regression Analysis Relating Second Grade Enrollments to High School Graduates

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.959[a]	.919	.911	201.779

a. Predictors: (Constant), grade2

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	4119.178	759.383		5.424	.000
	grade2	.603	.056	.959	10.686	.000

a. Dependent Variable: hsgrad

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4649217	1	4649217.152	114.190	.000[a]
	Residual	407148.8	10	40714.885		
	Total	5056366	11			

a. Dependent Variable: hsgrad

b. Predictors: (Constant), grade2

Application 1: High School Graduate Projections

The first application is for an institutional researcher who is asked by her director to derive projections of the numbers of high school graduates for their state. The analyst has data on the number of high school graduates each year ($Hsgrad_t$) for a 12-year period 1990-2001, and the second grade enrollments each year ($Grade2_t$) for a 22-year period 1980-2001. The data for this example are contained in the file HSGRAD.SAV. The model used by the analyst is that the number of high school graduates in any year will be a linear function of the number of second grade students ten years earlier, since second graders in a given year (such as 1985) will normally be eligible for high school graduation ten years later (such as 1995): $Hsgrad_t = \beta_0 + \beta_1 Grade2_{t-10} + e_t$.

The analyst would begin by estimating the equation shown above using data on second grade enrollments from 1980 to 1991 and high school graduates from 1990 to 2001. Note that by construction, the data set consists of only twelve observations and ten degrees of freedom. The SPSS output in Table 2 shows the results from the regression analysis.

The coefficient of determination is shown in the column with the heading

"R square." Despite the relatively small sample, the value of $R^2 = .919$ indicates that almost 92% of the annual variations in high school graduates for the state are explained by variations in the numbers of second grade students ten years earlier. In the second box, the total sum of squared deviations is broken down into the portions that are explained and not explained by the regression model. The fifth column in this table contains the F-statistic described earlier for assessing whether the regression equation as a whole has a significant impact on the dependent variable. Note that while the p-value for this test statistic is approximately zero (see the last column with the heading "Sig"), this is a fairly trivial test in practice and does not give any guidance as to whether or not the regression model is good.

Of more importance here are the results shown in the last output box. The column with the heading "B" under "Unstandardized Coefficients" contains the estimated coefficients for each of the variables in the model. Likewise, the estimated standard errors for each coefficient are in the column headed "Std. Error." The equation would be written as follows based on this output:

$$\hat{Hsgrad}_t = 4119.18 + .603 * Grade2_{t-10}$$

The calculated t-ratios for each of these variables are shown in the column headed "t." In this example, the calculated t-ratio = 10.69 and is highly significant, and thus the null hypotheses that second grade enrollments has no effect on high school graduates ten years into the future can easily be rejected. The results are encouraging even though the sample size on which the estimates were based was relatively small. These calculated t-ratios are obtained by dividing the estimated coefficients in column 2 by the standard errors in column 3. It is also worth noting that when the regression model contains only one independent variable, the F-statistic for the overall regression model is exactly equal to the square of the t-ratio for the independent variable.

In this application, recall that the analyst is interested not only in estimating the relationship between second grade enrollments and future high school graduates, but also using this model to predict the numbers of high school graduates in the future. This particular model is well suited to the task because the analyst has data on second grade enrollments from 1992-2001 that can be used to predict high school graduates for 2002-2011. Point estimates can be obtained by substituting second grade enrollments for each year into the estimated equation and solving for high school graduates. However, these predictions are likely to be inaccurate due to sampling variability, which is magnified here because of having only ten degrees of freedom. The analyst can also compute prediction intervals for each of these estimates. Doing so in this example would lead to the following predictions shown in Table 3.

These data show, for example, that in 2002 the model would predict that there will be 12,459 high school graduates in the state. Furthermore, the analyst is 95% certain that the number of high school graduates in 2002

Table 3
Point and Interval Predictions of the Numbers of High School Graduates based on Regression Model

Year	Predicted Numbers of High School Graduates		
	Lower Bound Estimate	Point Estimate	Upper Bound Estimate
2002	12,319	12,459	12,599
2003	12,778	12,986	13,193
2004	13,422	13,773	14,124
2005	13,736	14,163	14,590
2006	13,753	14,185	14,616
2007	13,709	14,130	14,551
2008	14,089	14,605	15,121
2009	14,004	14,499	14,993
2010	14,203	14,748	15,293
2011	14,339	14,919	15,498

will be between 12,319 and 12,599. This information would be useful to the institution in long-range planning for staffing, facilities, and recruiting.

Application 2: Faculty Salary Studies

The second institutional research application addresses how to measure the pay disparity between male and female faculty members at a university. The data for this example are contained in the file FACULTY.SAV and are stored as an SPSS system file. The following information is available on each of 432 faculty members:

Variable	Description
Rank	1=full professor, 2=associate professor, 3=assistant professor
Gender	1=males, 0=females
Nine12	1=9-month appointment, 0=otherwise
Cites	Number of citations received in a particular year
Annsal	Annual salary of faculty members
Yrsexp	Years of experience
Lsalary	Natural log of annual salary
Yrsexp2	Squared years of experience

Note that the variable for academic rank is a categorical variable. Accordingly, the first step is to create dummy variables for each of the three academic ranks (*Full* = 1 if *Rank* = 1, 0 otherwise; *Asso* = 1 if *Rank* = 2, 0 otherwise; *Asst* = 1 if *Rank* = 3, 0 otherwise). Descriptive statistics for these and selected variables are shown in Table 4.

The SPSS output shows that, on average, male faculty earn almost $11,000 more (or 20%) than female faculty. At the same time, on average, male faculty have received almost twice as many citations as females for

Table 4
SPSS Output: Descriptive Statistics for Faculty

Report

Mean

gender	annsal	yrsexp	full	asso	asst	cites
0	45841.14	14.7591	.2065	.4348	.3587	5.30
1	56833.92	19.9744	.5882	.3000	.1118	10.19
Total	54492.87	18.8637	.5069	.3287	.1644	9.15

their published work, have over five years of additional experience, and are more likely than females to be employed at the full professor level. Together, these factors might help account for the large difference in average salaries between male and female faculty.

To determine how years of experience, citations, and gender influence salary, the analyst might estimate a salary model (Model 1) and obtain the following results:

Table 5
SPSS Output: Regression Results for Faculty Salaries – Model 1

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.461[a]	.212	.207	13954.16965

a. Predictors: (Constant), yrsexp, cites, gender

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2.24E+10	3	7479066622	38.410	.000[a]
	Residual	8.33E+10	428	194718850.6		
	Total	1.06E+11	431			

a. Predictors: (Constant), yrsexp, cites, gender

b. Dependent Variable: annsal

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	38540.912	1854.690		20.780	.000
	gender	7689.289	1697.905	.201	4.529	.000
	yrsexp	408.273	76.692	.235	5.324	.000
	cites	240.277	36.824	.282	6.525	.000

a. Dependent Variable: annsal

The coefficient of determination is shown in the column with the heading "R square." The value of $R^2 = .212$ indicates that 21% of the variations in faculty salaries are explained by the variables gender, years of experience, and citations. In the second box, the p-value for the F-statistic is approximately zero, indicating that gender, experience and citations collectively have a significant impact on salary.

Using the output from the third box, the equation relating gender, experience and citations to salary can be written as follows:

$$\hat{Annsal} = 38540.92 + 7689.29 * Gender + 408.27 * Yrsexp + 240.28 * Cites$$

This model could be used to predict annual salary for a female professor with twenty years of experience and five citations by substituting her values for each variable into the equation:

$$\hat{Annsal} = 38540.92 + 7689.29 * (0) + 408.27 * (20) + 240.28 * (5) = \$47,907.72$$

The calculated t-ratios for each of these variables are shown in the column headed "t," and because each is highly significant, the null hypotheses that each variable has no effect on salary can easily be rejected. The coefficient on the variable *Gender* indicates that male faculty earn about $7,700 more than female faculty with the same number of citations and years of experience. Therefore, about 25% of the average salary difference is explained by these two factors.

Suppose now that the analyst wished to examine whether also controlling for academic rank would influence the findings of the study. To do this, the analyst adds two of the three dummy variables for rank into the salary model (Model 2) and obtains the following output from SPSS (see Table 6).

According to these results, 43% of the variation in salaries is explained by the addition of the two variables for current rank. The coefficient for the dummy variable *Full* indicates that Full Professors earn $22,989 more than their peers who have similar levels of experience, citations, and gender. Note, however, that the variable *Yrsexp* is no longer statistically significant. This dramatic change in sign and significance level is because of the high correlation between the rank and experience variables. This can be seen in the correlation matrix provided. Not surprisingly, a faculty member's years of experience is shown to be very highly correlated with whether the faculty member is a Full Professor. While the other variables are also correlated with gender, the correlation is too low to cause problems of multicollinearity. However, if there is gender discrimination in promotion at the institution, then controlling for rank in the salary model will lead to an underestimate of true pay disparity between men and women because some of the pay disparity is caused by slower promotion rates for women. It is at this point that the analyst should examine the model and determine which correlations with gender are appropriate and which correlations might be problematic due to gender bias.

Table 6
SPSS Output: Regression Results for Faculty Salaries – Model 2

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.656[a]	.430	.423	11897.25048

a. Predictors: (Constant), asso, cites, gender, yrsexp, full

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	37942.101	1814.660		20.909	.000
	gender	2959.464	1496.388	.077	1.978	.049
	yrsexp	-94.490	76.343	-.054	-1.238	.217
	cites	177.660	31.792	.208	5.588	.000
	full	22988.554	1962.131	.734	11.716	.000
	asso	8288.327	1782.094	.249	4.651	.000

a. Dependent Variable: annsal

Correlations

		yrsexp	full	asso	gender
yrsexp	Pearson Correlation	1	.521**	-.252**	.237**
	Sig. (2-tailed)	.	.000	.000	.000
	N	432	432	432	432
full	Pearson Correlation	.521**	1	-.710**	.313**
	Sig. (2-tailed)	.000	.	.000	.000
	N	432	432	432	432
asso	Pearson Correlation	-.252**	-.710**	1	-.117*
	Sig. (2-tailed)	.000	.000	.	.015
	N	432	432	432	432
gender	Pearson Correlation	.237**	.313**	-.117*	1
	Sig. (2-tailed)	.000	.000	.015	.
	N	432	432	432	432

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

In many universities, salary increases are given as a percentage of income rather than a fixed dollar amount. Accordingly, a semilogarithmic salary model may better represent the salary determination process at the institution. To test this idea, the analyst replaces the variable *Annsal* with *Lsalary* and reestimates the equation. The main results for Model 3 are shown in Table 7.

Table 7
SPSS Output: Regression Analysis Results for Faculty Salaries – Model 3

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	10.555	.029		361.010	.000
	gender	.065	.024	.098	2.679	.008
	yrsexp	-.003	.001	-.102	-2.466	.014
	cites	.003	.001	.199	5.692	.000
	full	.455	.032	.847	14.384	.000
	asso	.191	.029	.334	6.658	.000

a. Dependent Variable: lsalary

The resulting salary model is written as:

$$\widehat{Lsalary} = 10.555 + 0.065*Gender - .003*Yrsexp + .003*Cites + .455*Full + .191*Asso$$

The coefficient on the variable *Gender* suggests that after controlling for the effects of years of experience, citations, and current academic rank, male faculty earn approximately 6.5% more than female faculty. The level of significance for *Gender* is also notably higher than before, with a p-value = .008 compared to a p-value = .049 in the linear salary model.

Suppose now that the analyst wanted to determine if there is a quadratic relationship between a faculty member's actual salary and his/her level of experience. To test this hypothesis, the variable for squared experience (*Yrsexp2*) is added to the regression model, and the following results (Model 4) are obtained:

Table 8
SPSS Output: Regression Analysis Results for F aculty Salaries – Model 4

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	34377.041	2933.688		11.718	.000
	gender	7421.761	1699.584	.194	4.367	.000
	cites	242.275	36.740	.284	6.594	.000
	yrsexp	940.477	300.934	.542	3.125	.002
	yrsexp2	-13.004	7.112	-.315	-1.829	.068

a. Dependent Variable: annsal

The effect of experience on salary (ignoring the other variables in the model) can be written as:

$$\hat{Annsal} = 940.48 * Yrsexp - 13.00 * Yrsexp2$$

The fact that the estimated coefficient on the squared variable is negative indicates that the quadratic curve is hill-shaped and not U-shaped. The change in salary due to a one-year increase in experience is found by taking the first partial derivative of this equation:

$$\text{Change in } Annsal \text{ due to change in } Yrsexp = 940.48 - 26 * Yrsexp$$

Accordingly, annual salary increases at a decreasing rate as a faculty member's years of experience rises up to 36 years of experience, and then would begin to increase at an increasing rate.

Summary

This chapter has provided an overview of regression analysis and how it can be used by institutional researchers. Regression analysis is an attractive option for statistical inquiries because of its flexibility. The technique can be used for many types of variables – continuous, discrete, dichotomous – and through variable transformations can accommodate even categorical variables into the analysis. Likewise, appropriate variable transformations can enable regression models to estimate non-linear relationships between variables of interest. Regression analysis is valuable for not only testing hypotheses about whether specific factors are related to each other, but also for quantifying the relationships and using these to derive forecasts of dependent variables of interest.

Suggested Readings

Bellas, M., & Toutkoushian, R. (1999). Faculty time allocations and research productivity: Gender, race, and family effects. *The Review of Higher Education 22*(4), 367-390.

Betts, J., & Grogger, J. (2003). The impact of grading standards on student achievement, educational attainment, and entry-level earnings. *Economics of Education Review 22*(4), 343-352.

Gujarati, D. (1995). *Basic Econometrics*, 3rd Edition. New York: McGraw-Hill, Inc.

Hsing, Y., & Chang, H. (1996). Testing increasing sensitivity of enrollment at private institutions to tuition and other costs. *The American Economist 40*(1), 40-45.

Mincer, J. (1974). *Schooling, Experience, and Earnings*. New York and London: Columbia University Press.

Pennock-Roman, M. (2002). Relative effects of English proficiency on general admissions tests versus subject tests. *Research in Higher Education 43*(5), 601-623.

Powers, J. (2004). R&D funding sources and university technology transfer: What is stimulating universities to be more entrepreneurial? *Research in Higher Education 45*(1), 1-24.

Sax, L., Hagedorn, L., Arrendondo, M., & Dicrisi, F. (2002). Faculty research productivity: Exploring the role of gender and family-related factors. *Research in Higher Education 43*(4), 423-446.

Schonwetter, D., Clifton, R., & Perry, R. (2002). Content familiarity: Differential impact of effective teaching on student achievement outcomes. *Research in Higher Education 43*(6), 625-656.

Toutkoushian, R. (1998). Using regression analysis to determine if faculty salaries are overly compressed. *Research in Higher Education 39*(1), 87-100.

Toutkoushian, R. (1999). The value of cost functions for policymaking and institutional research. *Research in Higher Education 40*(1), 1-16.

Toutkoushian, R., Dundar, H., & Becker, W. (1998). The National Research Council graduate program ratings: What are they measuring? *The Review of Higher Education 21*(4), 427-443.

Toutkoushian, R., & Smart, J. (2001). Do institutional characteristics affect student gains from college? *The Review of Higher Education 25*(1), 39-61.

Venti, S., & Wise, D. (1983). Individual attributes and self-selection of higher education: College attendance versus college completion. *Journal of Public Economics 21*, 1-32.

Volkwein, J., & Zhou, Y. (2003). Testing a model of administrative job satisfaction. *Research in Higher Education 44*(2), 149-172.

Wooldridge, J. (2000). *Introductory Econometrics*. Cincinnati, OH: South-Western Publishing Co

Chapter 4
What Can Multilevel Models Add to Institutional Research?

Stephen Porter

Why should institutional researchers care about multilevel modeling techniques? Because institutional research is, at its heart, the analysis of institutional structures, and multilevel models offer one of the best ways to accurately understand the effects of these structures on faculty and students. We are often interested in assessing the impact of structures within our institutions, such as academic departments, or more often, in describing and understanding differences between institutions. For example, much of our data collection efforts, such as the IPEDS surveys and college guidebook requests, are used to compare and contrast institutions. It is only natural that we use statistical techniques that can appropriately analyze the complexity of our institutions, and in turn, the complexity within our data.

The advantage of multilevel models lies in their analytical approach. Analyses of survey data collected from students across multiple colleges generally use traditional multiple regression, or ordinary least squares (OLS), to understand individual and institutional correlates of student attitudes and behavior (e.g., Hu & Kuh, 2003a; Kuh & Hu, 2001; Toutkoushian & Smart, 2001). With this approach one model is estimated, and the coefficients of the model are constrained to be constant across schools. Thus, the impact of being female on engagement, for example, is estimated to be the same for each school. Multilevel models (or as they are also called, hierarchical linear models), estimate a model for each school in the sample, even for schools with few student observations. Often the values of these coefficients will differ for each school, so there may be no difference in engagement between males and females in some schools, while in other schools the difference may be quite large. This variation is of substantive interest, and we can use variation in institutional structures to analyze variation in these coefficients. What aspects of colleges, for example, are successful in ameliorating differences in engagement between males and females, and conversely, what college structures exacerbate these differences? Multilevel models can answer these questions in ways that OLS cannot.

The difference between multilevel models and OLS is more than a statistical quibble. Because regular regression techniques do not take into account the grouping of individuals within organizations, they can yield biased coefficients and standard errors when analyzing these data. This means that we may draw the wrong conclusions about how college structures affect individuals. Are liberal arts colleges better at creating diverse learning environments (Umbach & Kuh, in press)? Do students at research universities experience less contact with faculty than students at other types of institutions

(Kuh & Hu, 2001)? Are there differences in student development between historically Black institutions and historically White institutions (Kim, 2002b)? These are important questions, and as educational researchers we want to minimize the chance that our answers to these questions are misleading.

While there are a large variety of multilevel models, I focus here on an application most likely to be of interest to institutional researchers: the analysis of individuals nested within academic organizations such as departments or colleges. Most of the recent multilevel research in higher education has focused on analyzing students and faculty within multiple departments and multiple colleges, so this approach will also provide the reader with the background to understand this common application of multilevel models. The end of this chapter will briefly describe other applications of multilevel models.

After first discussing the advantages of multilevel models over OLS, I will review the theoretical background of these models. The third section presents an overview of practical modeling considerations, and the fourth section uses the 1998 Beginning Postsecondary Student Survey to illustrate the multilevel approach. The chapter concludes by reviewing further reading and software.

OLS versus Multilevel Models

Three issues arise when using OLS to analyze data on individuals grouped within organizations: misestimated standard errors, aggregation bias, and heterogeneity of regression coefficients (Raudenbush & Bryk, 2002, pp.99-100).

First, a fundamental assumption of OLS is that the error terms are not correlated across observations. Simply put, "this means that in repeated samples there is no tendency for the disturbance associated with one observation (corresponding, for example, to one time period or one individual) to be related to the disturbance associated with any other" (Kennedy, 2003, p. 134). This assumption is likely to be violated when individuals undergo similar experiences, such as students learning within colleges, or faculty working within academic disciplines. The result is misestimated standard errors for individual-level variables; therefore, hypothesis tests for these variables may be faulty.

Additionally, we often want to estimate the impact of departmental or college attributes; for example, is student satisfaction higher in schools with low student-faculty ratios? If we attach school-level data to our individual-level data and run a standard regression analysis, the standard errors for the school-level variables will be underestimated. OLS assumes the number of schools is the same as the number of individuals, when in fact the number of school-level observations is much smaller.

Second, we may be interested in disentangling person-level and compositional effects. Suppose we aggregate our student data to the college level, so that we now have average SAT scores for each school. If we use

these to predict graduation rates and find a relationship, what can we conclude? The issue here is that SAT score is both a measure of individual student aptitude as well as a measure of institutional selectivity, two related but theoretically distinct concepts. Multilevel models can appropriately disentangle these effects, so that we can understand the impact of a student's academic background, as well as the impact of attending a selective institution.

Third, OLS cannot handle substantial heterogeneity of regression coefficients; that is, regression coefficients that differ for each group in a sample. Instead, regression coefficients are fixed across groups. Multilevel models can estimate a different coefficient for each group; for example, the impact of socioeconomic status on student satisfaction can differ for each college in a multi-college data set. More importantly, the reason why the impact differs can in turn be explained by the multilevel model. The classic example is the moderating effect of parochial schools on the relationship between a secondary student's socioeconomic status and their math achievement (Raudenbush & Bryk, 2002, pp. 119-130).

Theoretical Background

Many people new to multilevel models have difficulty in understanding the approach, not because of the mathematics, but because they have been taught to consider regression coefficients to be fixed. Conversely, in the multilevel approach these coefficients can vary. One of the easiest ways to understand the multilevel approach is to consider the "slopes as outcomes" approach.[1]

Suppose we have an engagement survey of college students from fifty different colleges, with 100 responses for each school, so that the total survey N equals 5,000 students. Rather than run an OLS model on the entire sample, we could instead run a separate regression model for each school, analyzing how student attributes such as gender and socioeconomic status affect student engagement within each school. We could then take the results for each school and enter the intercepts and coefficients for the independent variables into a spreadsheet. This would give us fifty lines of data, one for each school regression model. By adding additional columns of data that describe the schools, such as number of undergraduates and expenditures per student, we could run a second-stage regression that would explain variation in the schools' intercepts and coefficients due to student body size and level of resources. For the types of multilevel models used to understand the impact of organizations, this is, in essence, how multilevel models work, albeit with different statistical techniques than OLS.

We can describe how these models work more formally with equations. Use of these equations is necessary for two reasons. First, the models can become quite complex as both the intercept and coefficients are allowed to vary; using simple algebraic notation allows us to keep track of and understand exactly what we are estimating. Second, understanding the notation is crucial

to using and interpreting the software programs that estimate multilevel models. The software program HLM, for example, uses this notation to present results, while the user must use equations to determine the structure of the program code in SAS PROC MIXED (Singer, 1998).

To review multilevel notation, I use the classic example of student high school math achievement as predicted by socioeconomic status (SES) and high school public/private status. This is the example used by Raudenbush and Bryk (2002, chapter 4) and other researchers, and will make reading their texts easier if the reader wishes to learn more about multilevel models.

The Random Intercept Model

The random intercept model is probably the most common multilevel model seen in higher education literature, and is used to correctly estimate the impact of institutional structures on individual behavior. The name stems from the single random component of the model, the intercept: a separate intercept is estimated for each school, but the regression coefficients for the independent variables are held constant across schools. Recent examples include Hu and Kuh (2002), Kim (2002b), and Umbach and Porter (2002).

Equation 4.1 describes the familiar OLS regression equation, where Y_i is the math achievement test score of student *i*. As with any regression equation, the intercept, β_0 is the expected value of the dependent variable when *X* (SES) equals zero, while the regression coefficient β_1 is the expected change in math achievement given a one unit change in student SES. Each student has a random error term, r_i which reflects the difference between the student's predicted math achievement and their actual math achievement test score. These errors are assumed to be normally distributed with a mean of zero and variance σ^2.

$$Y_i = B_0 + B_1 X_i + r_i \qquad [4.1]$$

For both interpretive and mathematical considerations, the independent variables in multilevel models are often rescaled by subtracting the mean value of the independent variable $(\bar{X})$ from each observation. If the mean is the overall mean of the sample, such rescaling is referred to as *grand-mean centering*; SES can now be represented as $(X_i - \bar{X})$ so that Equation 4.1 can be rewritten as:

$$Y_i = B_0 + B_1 (X_i - \bar{X}) + r_i \qquad [4.2]$$

Table 1 shows several hypothetical students and their SES backgrounds. In this example, SES is measured on a scale that ranges from one to ten, and the mean SES for the entire sample is four. Note that with grand-mean centering, students with SES above the mean have positive scores, students below the mean have negative scores, and students at the mean have a

score of zero. Thus grand-mean centering provides a more meaningful interpretation of the intercept β_0: rather than the math achievement of a student with zero SES (a nonsensical interpretation given that the original SES scale ranges from one to ten), β_0 now represents the math achievement of a student with average SES. Other values can also be used for centering, such as the group means or values determined by substantive theory.

Table 1
Rescaling with Grand-Mean Centering

Student	SES score (1-10 scale)	
	Raw	Grand mean centered ($\bar{X} = 4$)
A	2	-2
B	7	3
C	1	-3
D	4	0
E	10	6

Equations 4.1 and 4.2 are the standard regression equations found in any basic econometrics textbook. Note that when applied to student data from multiple institutions, the effect of SES (B_1) is the same for every school, and average SES (B_0) is also the same for every school. But Equation 4.2 can be rewritten to represent separate equations for each school:

$$Y_{ij} = B_{0j} + B_1(X_{ij} - \bar{X}_{..}) + r_{ij} \qquad [4.3]$$

Here, the *j* subscripts indicate individual schools, just as the *i* subscripts indicate individual students. Y_{ij}, X_{ij}, and r_{ij} have j subscripts to indicate the scores and error terms for student *i* in a particular school *j*, and the two period subscripts for $\bar{X}$ indicate that this mean has been calculated over the entire sample and is not conditional upon schools or individuals. Because we are now estimating a separate equation for each school, each B_0 has a *j* subscript to indicate the intercept for school *j*, but the effect of SES remains constant across schools.

We can use Equation 4.3 to describe an individual student's math score as comprised of three parts. Each student's score is determined by the average math achievement score for their school j (B_{0j}), a deviation from this average based on B_1 times the student's SES score (with no deviation for the student with an average SES score), and a unique student deviation or error term r_{ij}. Equation 4.3 is referred to as the *level 1* model; it explains the variance in the dependent variable at the first level of our data, the student level.

Because B_{0j} varies across schools, we can model this variance using a second equation at *level 2*, the school level or second level of our data.[2] The simplest level 2 equation explains the variance in the school intercepts as follows:

$$B_{0j} = \gamma_{00} + u_{0j} \qquad [4.4]$$

Here B_{0j} is a function of γ_{00} the average of the school means for math achievement, and a unique school-level deviation or error term u_{0j}. These errors are assumed to be normally distributed with a mean of zero and variance τ_{00}. (The subscripts in multilevel models can be somewhat confusing. At level 2, the first subscripted number indicates variables and coefficients associated with a level 1 coefficient; here the intercept at level 1, B_{0j} has a subscript 0, so all level 2 variables and coefficients explaining the variance in B_{0j} will have a 0 as the first subscripted number. The second subscripted number refers to the variable order in the level 2 equation; here there is only one variable, but as variables are added the second subscripted number will increase.)

The level 2 equation can be more complex. Suppose we believe that math achievement varies between schools in large part because of their public/private status. We can add a dummy variable W to our model that takes a value of one for private schools, zero for public schools. The level 2 equation contains this school-level variable to explain the variance in the school intercepts:

$$B_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j} \qquad [4.5]$$

Note that interpretation changes here with the addition of explanatory variables at level 2. Inclusion of a dummy variable changes the interpretation of the intercept, just as in OLS. B_{0j} is now a function of γ_{00}, the mean of the dependent variable for private schools, W_j, the public-private dummy variable, γ_{01} a coefficient measuring the impact of private school status on B_{0j}, and a school-level error term u_{0j}. Additional school-level variables could also be included in this equation, such as school size or racial makeup of the student body.

In terms of substantive theory, this statistical approach is similar to the equations estimated by Toutkoushian and Smart (2001), who used variables such as student body size and expenditures per student, as well as student-level variables, to understand the impact of institutions on student gains in college. While they used OLS, the goal of their model is the same as that of Equations 4.2 and 4.3: explain the variation in student gains with student variables such as gender and race, and college variables such as size and finances.

In sum, the random coefficient model is substantively similar to the traditional regression model run on combined student and school data. The level 1 model estimates the impact of student-level variables on the dependent variable, and produces a set of intercepts for each school. These intercepts have been adjusted for differences in the makeup of the student body across schools due to the inclusion of the level 1 variables. Variation in these intercepts is then explained by school-level variables in the level 2 model.

The Random Coefficient Model

The only difference between the random intercept model and random coefficient model is that one or more coefficients for the independent variables have been allowed to vary across schools, so that a different coefficient is estimated for each school. It is less common in the higher education literature, in part because estimating additional random components beyond the intercept requires considerable data. Examples include Hu and Kuh (2003b), Porter and Umbach (2001), and Rumberger and Scott (1993).

This difference is illustrated in Equation 4.6, where the slope coefficient B_1 now has an additional subscript *j*, indicating that it also varies across schools. Substantively, this means that the impact of SES on math achievement is now different for each school. In some schools there may be no difference in math achievement between high and low SES students; in these schools, B_1 is very close to zero. In other schools, B_1 may be large and positive, indicating that high SES students have higher math achievement scores than low SES students.

$$Y_{ij} = B_{0j} + B_{1j}(X_{ij} - \bar{X}_{.j}) + r_{ij} \qquad [4.6]$$

This small change now allows us to investigate a very interesting substantive question: what is it about some schools that makes it possible for them to erase performance differences between high and low SES students? Suppose our hypothesis is that private schools are more egalitarian in their outcomes; that is, we think that the impact of SES is much smaller for private than public schools. We can test this hypothesis with an additional level 2 model.

Before constructing the level 2 model, we must first revisit the issue of centering. Centering is one of the more contentious areas in multilevel modeling, because reasonable scholars can disagree as to whether a variable should be centered, and if so, how. For the random intercept model, most analyses will either not center the independent variables or use grand-mean centering. For the random coefficient model, described here to investigate organizational effects, we can rely on a simple rule of thumb: when the coefficient of an independent variable is randomized, that is, when we run a level-2 equation explaining variance in the coefficient, then the independent variable will be *group-mean centered.* As with any rule of thumb there will be exceptions, but in most analyses this will be the appropriate centering choice.

Group-mean centering indicates that the independent variable is rescaled by subtracting from each observation the mean value of the independent variable *for the group* rather than the mean value for the entire sample. Table 2 revisits the hypothetical students and their SES scores, which are measured on a scale that ranges from one to ten; students are grouped in three different

schools. Each student's SES is now represented by $(X_i - \bar{X}_{.j})$ that is, the mean SES for their school *j* is subtracted from their SES score. For example, student F in school 2 has a relatively high SES score, 7. Because the other students in school 2 also have high SES scores (the school mean is 6.67), student F's group-mean centered SES score is now .33, indicating that student F is .33 SES units above the average student in school 2.

Table 2
Rescaling with Group-Mean Centering

		SES score (1-10 scale)	
School	Student	Raw	Grand mean centered
1	A	2	-2.80
1	B	7	2.20
1	C	1	-3.80
1	D	4	-.80
1	E	10	5.20
$\bar{X}_{.1}$		**4.80**	**.00**
2	F	7	.33
2	G	8	1.33
2	H	5	-1.67
$\bar{X}_{.2}$		**6.67**	**.00**
3	I	3	-.75
3	J	6	2.25
3	K	4	.25
3	L	2	-1.75
$\bar{X}_{.3}$		**3.75**	**.00**

In this application group-mean centering is preferred for estimation reasons; however, by group-mean centering SES, we have lost the information that differentiates the schools in terms of SES. In Table 2, students F and K have very similar group-mean centered SES scores (.33 and .25), but their raw SES scores differ quite a bit (7 and 4), because the mean SES for the two schools differs (6.67 versus 3.75). These group means are usually introduced back into the model in the level 2 equations.

Recall that our first level 2 model explained variation in the intercepts with schools' public/private status. We can construct a similar model for the SES coefficient:

$$B_{1j} = \gamma_{10} + \gamma_{11} W_j + \gamma_{12} \bar{X}_{.j} + u_{1j} \qquad [4.7]$$

Here B_{1j} is now a function of γ_{10}, the average effect of SES on math achievement for private schools, W_j, the public-private dummy variable, γ_{11}, a coefficient measuring the average effect of SES for private schools, γ_{12}, the effect of mean school SES on the impact of SES at the individual level, and a school-level error term u_{1j}. As with the intercepts, the errors u_{1j} are assumed to be normally distributed with a mean of zero and variance τ_{11}.

To test our hypothesis about private schools, we would estimate a multilevel model using Equations 4.6 and 4.7 and an equation similar to Equation 4.7 for the intercepts (as explained below, when estimating a level 2 model for coefficients it is advisable to also estimate the same model for the intercepts). We would expect both γ_{10} and γ_{11} to be statistically significant, γ_{10} to be large and positive, indicating that high SES students have higher math achievement than low SES students in public schools, and γ_{11} to be negative, indicating that the impact of SES is lower in private schools compared with public schools.

If γ_{12} were statistically significant and negative, this would indicate that the differences in math achievement between high and low SES students (recall that this is the interpretation of the level 1 coefficient, B_{1j}) are reduced as the mean SES for a school increases. If we consider mean student SES for each school $(\bar{X}_{.j})$ as a proxy for school and neighborhood financial resources, a negative value for γ_{12} indicates that the values of the level 1 coefficient for SES decreases as schools become wealthier: wealthy schools are more egalitarian. The reverse would be true if γ_{12} were positive.

We can see that random coefficient models can become exceedingly complex, with possibly a separate level 2 equation for each level 1 variable. In practice, the data usually cannot handle so many different random effects and equations, and often only one or two level 1 coefficients are randomized. For example, Hu and Kuh (2003b) estimate a level 1 model where student gains in growth and development are predicted by student effort and other student-level variables, while the intercept is randomized and a level 2 model including variables such as selectivity and Carnegie type is estimated. In addition, they randomize one regression coefficient, student effort, and estimate the same level 2 model for the student effort regression coefficient. Thus, they investigate not only the impact of institutional characteristics on student gains, but they also explore why some institutions are successful in translating student effort into academic development and others are not.

Summary

Figure 1 shows the full sets of equations for the random intercept and random coefficient models. With the random intercept model, a student-level equation is estimated for each school, where the intercept differs for each school but the impacts of the independent variables are constrained to be equal across schools. With the random coefficient model, this constraint is relaxed and a different coefficient is estimated for each school. Variation in these intercepts and coefficients is explained by school-level variables in the level 2 models. For statistical reasons, in random coefficient models, the level 1 variables are usually group-mean centered and the group means are used as independent variables at level 2; in random intercept models the variables are either left uncentered or are grand-mean centered.

Applied Modeling Considerations

This section reviews some of the practical aspects of multilevel modeling, beginning with the first question that should be answered: should one use multilevel modeling for an analysis? The answer depends on the type of data and the variation within the data.

Figure 1
Two Approaches to Understanding Organizational Effects

Random intercept model

Objective: understand the impact of group-level variables on the dependent variable
Centering: usually uncentered or grand-mean

Level 1 $Y_{ij} = B_{0j} + B_1(X_{ij} - \bar{X}_{..}) + r_{ij}$ Explain variation in dependent variable

Level 2 $B_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j}$ Explain variation in intercepts

Random coefficient model

Objective: understand how the group-level variables explain variation in the individual-level regression coefficients
Centering: usually group-mean

Level 1 $Y_{ij} = B_{0j} + B_1(X_{ij} - \bar{X}_{.j}) + r_{ij}$ Explain variation in dependent variable

Level 2 $B_{1j} = \gamma_{10} + \gamma_{11}W_j + \gamma_{12}\bar{X}_{.j} + u_{1j}$ Explain variation in slopes

Level 2 $B_{0j} = \gamma_{00} + \gamma_{01}W_j + \gamma_{02}\bar{X}_{.j} + u_{0j}$ Explain variation in intercepts

Data Requirements

Whether multilevel modeling can be used depends on the structure of the data. In general, most multilevel modelers recommend a minimum of thirty groups, with the number of individuals per group averaging at least ten (note that it is possible to have a few groups with only one individual). This is only a rule of thumb, and smaller sample sizes can be used, but consider why we would want large numbers of individuals within groups and large numbers of groups.

First, multilevel models estimate coefficients for each group. These coefficients are subject to sampling error, as they are derived from samples from each group; for example, we may have only twenty students from a school of 3,000 students. As the sample size for a group increases, the reliabilities for the group estimate become larger. HLM estimates are based

on a weighted average of the group estimate and the entire sample mean; as reliabilities increase, more weight is placed on the group mean. Thus, having more individuals per group rather than less, generally results in a larger variance in group-level estimates.

Second, the level 2 models are estimated using the group coefficients and group characteristics as data; the N at level 2 is much smaller than at level 1. Thus the greater the number of groups, the smaller our standard errors will be. In addition, more groups will often permit more randomized coefficients.

For discussions of the number of groups and group sizes needed, see Raudenbush and Bryk (2002, chapter 9) and Heck and Thomas (2000, pp. 26-30). These discussions are especially helpful when planning a research study, as one can use power analyses to derive estimates of the number of groups and group sizes needed.

Intraclass Correlation: Proportion of Variance between Groups

Whether multilevel modeling should be used is contingent on how much variation in the dependent variable is explained by group membership. It is calculated by running the *null model*, a multilevel model with no variables at either level 1 or level 2, so that only the individual level variance and group-level variance components are estimated. The *intraclass correlation* (ICC) is a measure of the variation in the dependent variable between groups and is given by the following formula:

$$\rho = \frac{\tau_{00}}{\tau_{00} + \sigma^2} \qquad [4.8]$$

where σ^2 is the measure of the within group or individual-level variance, and τ_{00} is the between group variance, or the variance in the intercepts. As can be seen, as the variance in the intercepts τ_{00} grows larger, ρ grows larger, and if τ_{00} is close to zero, the ICC will be close to zero.

The often quoted rule of thumb is that multilevel modeling is appropriate if the ICC is greater than .05; that is, if at least 5% of the variance in the dependent variable is between groups. However, multilevel models with statistically significant variables at level 2 can be estimated even when the ICC is low as .4% (see e.g., Merlo et al., 2001). This rule of thumb should be reinterpreted. One should not be surprised if multilevel modeling results are very similar to OLS results when the ICC is less than 5%, as this indicates that the group means for the dependent variable are very similar; hence there is little variation to model between groups. This variation can still be successfully modeled, and multilevel models should generally be used on such data if group-level variables are included in the model.

In general, the ICC will vary between zero and .40 in most social science research (Snijders & Bosker, 1999, p. 151). As an illustration of the typical ICC in higher education, the ICCs reported in several higher education research articles are listed in Table 3. Most of the ICCs range between 5% to 10%,

Table 3
Intraclass Correlations in Higher Education Multilevel Research

Study	Data	Dependent variables	ICC
Hu and Kuh (2003b)	44,238 students in 120 schools	7 self-rated learning scales	.03 to .09
Johnsrud & Rosser (2002)	1,511 faculty in 10 schools	6 work related scales	.01 to .15
Kim (2002b)	1,069 students in 81 schools	3 self-rated ability scales	.15 to .25
Kim (2002a)	1,397 students in 86 schools	3 self-rated ability scales	.16
Porter & Umbach (2001)	1,104 faculty in 103 academic disciplines	Publications and grants	.18, .32
Rosser et al. (2003)	865 faculty and staff rating 22 deans and directors	7 evaluation scales	.07 to .16
Rumberger & Thomas (1993)	7,235 students in 36-146 schools, depending on major	Salaries of college graduates for 6 major groups	.12 to .25
Smyth & McArdle (2004)	5,047 students in 23 schools	Science major graduation from college	.04
Strauss & Volkwein (2004)	8,217 students in 51 schools	Institutional commitment scale	.10
Thomas (2000)	3,382 students in 328 schools	Earnings and debts of college graduates	.08, .10
Umbach & Porter (2002)	1,532 students in 54 academic departments	4 satisfaction and self-rated development scales	.06 to .08

with more objective data, such as salaries or publications, having higher ICCs than more subjective data, such as scales based on self-ratings.

Building Models and Randomizing Coefficients

Once the decision is made to use a multilevel modeling approach, the first step is building the random intercept model. This involves completely specifying the individual-level model first, and then building the group-level model. In terms of higher education, we can think of this as first building a model explaining why our dependent variable varies within colleges, and then having explained the within-college variation, next building a model that explains variation between colleges.

The next step is deciding whether to estimate a random coefficient model; that is, to explain why some of the coefficients for the independent variables differ between colleges. The actual specification of the random coefficient model can be complex, because a model can be estimated for every independent variable in the model (although the data may not support such a complex model). Here, both theory and statistics will guide the model building.

When determining which coefficients should be random, theory should drive the choice. For example, natural choices for randomization are the coefficients for gender and race. Voluminous research has documented differences in student outcomes between males and females and between

racial and ethnic groups, and it is possible that these differences may vary between institutions because of differences in institutional structures; for example, HBCUs and women's colleges may provide different outcomes for Blacks and women than other institutions (Kim, 2002a, 2002b).

If we have theoretical reasons why we expect that the effect of a variable may differ between colleges, we can use a statistical test to determine if this is indeed the case. Software programs that estimate multilevel models produce not only the variance components for the intercepts and coefficients that have been randomized, but they also test the null hypothesis that each variance component is zero. If we can reject the null hypothesis, that is, if the variance component for a randomized coefficient is statistically significant, we can conclude that a particular regression coefficient does indeed vary between schools.

Finally, if a random coefficient model can be used, the issue arises concerning what variables should be used to explain variation in a coefficient. In general, most researchers recommend similar models for the intercept and slope(s). Often, the randomized intercepts and slope coefficients may be correlated. If group-level variables are used in the slope model and not in the intercept model, due to the intercorrelation these variables may show an effect on the slopes even if the group-level variables only explain variance in the intercepts.

Measures of Variance Explained

Similar to OLS, variance explained measures can be calculated for multilevel models, but because there is both an individual-level model and a group-level model, there are two variance explained measures in a two-level multilevel model. Recall that the null model is estimated by running a multilevel model with the intercept randomized and no additional independent variables. Two variance components are estimated: the individual variance σ^2 and the variance in the intercepts (or school-level means) τ_{00}. By comparing the variance components from our full model (the model with one or more independent variables) to the variance components from the model with no independent variables, we can calculate how much variation in our dependent variable at levels 1 and 2 is explained by the full model.

Suppose we estimate a random intercept model with several independent variables at both the individual and school level; we can calculate the variance explained using the following formulas:

$$\text{Variance explained at level 1} = \frac{\sigma^2_{\text{null model}} - \sigma^2_{\text{full model}}}{\sigma^2_{\text{null model}}} \qquad [4.9]$$

$$\text{Variance explained at level 2} = \frac{\tau_{\text{null model}} - \tau_{\text{full model}}}{\tau_{\text{null model}}} \qquad [4.10]$$

With both formulas, we can see that if the variance components from the full model are similar to the null model, then the variance explained is zero, and as the variance components from the full model become smaller

(because some of the variation is being explained by the independent variables), the variance explained statistics become larger. In addition, similar to the R-square, these measures will generally fall between zero and one.

Case Study: Student Engagement Across Institutions

To illustrate the multilevel approach, I use the Beginning Postsecondary Student Survey, a panel study of college students conducted by the National Center for Education Statistics beginning in the 1995-1996 academic year. I include only students enrolled in a Carnegie Research, Doctoral, Comprehensive or Liberal Arts College, resulting in 4,481 students in 360 schools with an average of twelve students per school. In other words, the level 1 units are students, and the level 2 units are colleges and universities. From the survey I constructed a factor score of student engagement, using nine questions that ask the students how often they engaged in activities such as meeting with faculty, writing papers or using the library (alpha=.67).

Two sets of independent variables are used to predict student engagement. At the student level, four dummy variables measuring whether a student is *female*, *non-white*, a *first-generation* college student, or a *full-time* student are included, as well as two continuous variables, *SAT score* and first-year *college grade-point average* (GPA). At the school level, several variables are included to control for major differences between institutions. *Enrollment* and a *squared enrollment* term to allow for nonlinearities control for differences in size, while Barron's *selectivity* index controls for differences in selectivity (this measure varies from zero for non-competitive schools to five for the most competitive schools). The *percentage of female undergraduates* controls for differences in the student body, and the *percentage of graduate students* in the total student body is included as a measure of institutional emphasis on research. Finally, two dummy variables for *Historically Black Colleges and Universities* (HBCU) and *public* institutions are also included in the model as control variables.

Table 4 presents results for the multilevel random intercept models and corresponding OLS models. In all models in this table, the student-level variables are grand-mean centered, the school-level continuous variables are grand-mean centered, and the school-level dummy variables are not centered.

Model 1 is the null model, a multilevel model with no independent variables and only the intercept randomized. From the variance components at the bottom of the table, we can see that the ICC equals .14 (.1383/(.8161+.1383)), so about 14% of the variation in engagement scores is between colleges. Note that the intercept is close to zero and not statistically significant. The intercept shows the average value of the dependent variable, and because the dependent variable is a factor score with mean zero, this result is exactly what we would expect to see.

Table 4
Correlates of Engagement, Multilevel and OLS Estimation

	Multilevel random intercept model			OLS	
	1	2	3	4	5
Student-level (N=4,481)					
Female		.1390**	.1327**	.1612**	.1390**
		(4.81)	(4.49)	(5.47)	(4.58)
Non-white		.1652**	.1593**	.1401**	.1441**
		(4.42)	(4.22)	(4.09)	(4.02)
1st generation		-.0873**	-.0757*	-.1306**	-.0941**
		(-2.88)	(-2.51)	(-4.29)	(-3.15)
SAT score		-.0006**	-.0007**	-.0005**	-.0007**
		(-6.45)	(-7.50)	(-5.68)	(-7.82)
College GPA		.0012**	.0011**	.0012**	.0011**
		(6.39)	(6.11)	(6.37	(5.70)
Full-time		.2290**	.2057**	.2854**	.2240**
		(5.27)	(4.79)	(6.64)	(5.32)
School-level (N=360)					
Enrollment			-.0167**		-.0166**
			(-2.61)		(-3.96)
Enrollment squared			.0002*		.0002*
			(2.01)		(3.12)
Selectivity			.0975**		.0921**
			(4.19)		(5.72)
% female students			.1106		.0994
			(0.94)		(1.11)
% graduate students			-.0920		-.0805
			(-0.45)		(-0.56)
HBCU			-.0895		-.0840
			(-0.81		(-1.08)
Public			-.2714**		-.2847**
			(-4.20)		(-6.45)
Intercept	.0362	.0236	.2041**	.0187	.2205**
	(1.42)	(0.94)	(4.51)	(1.30)	(7.09)
Variance components					
σ^2	.8161**	.7901**	.7915**		
τ_{00}	.1383**	.1328**	.0750**		
ICC	.14				
Student-level: % var. exp.		.03	.03		
School-level: % var. exp.		.04	.46		
adj. R-square				.04	.09

Note: t-statistics are shown in parentheses; p<.01 **, p<.05 *. p<.10 +.
All variables are grand-mean centered except for HBCU and public.

Model 2 estimates a model of engagement using only student-level (or level 1) variables; the corresponding OLS model is shown in Model 4. In general, the results are similar for the two models, although there are some differences. The coefficient for first-generation college student status in the multilevel model, for example, is about a third smaller than the corresponding coefficient in the OLS model. Both models indicate that females, non-whites, second-generation and full-time students have higher levels of engagement than male, White, first-generation and part-time students. Although the coefficients for SAT score and GPA are statistically significant, (which is not surprising given the student N of over 4,000 students), their substantive impact is almost zero.

Models 3 and 5 show the full multilevel and OLS model results when the school-level (or level 2) variables are included in the models. Comparing the coefficients across the two models, the results appear similar, but a comparison of the t-statistics (in parentheses) shows large differences between the two approaches. The t-statistics for the school-level variables are much larger in the OLS model than the multilevel model; for the enrollment variable the difference is 2.6 versus almost 4, for the public school dummy variable the difference is 4.2 versus 6.5. The difference in t-statistic values indicates one of the biggest problems in using OLS to estimate institutional effects on individual-level data: the statistical significance of the institutional variables will be overstated.

The variance components for Model 3 are shown at the bottom of the table, and using Equations 4.9 and 4.10 we can determine the amount of variation explained by the model. The variance explained at the student level is rather small, 3% ((.8161-.7915)/.8161), while the variance explained at the school level is substantial, 46% ((.1383-.0750)/.1383). Clearly the student-level or within-college model could be better specified in this example.

Comparing the intercept in Model 3 to the intercepts in Models 1 and 2, we can see that the intercept is now positive and statistically significant. Because the value of the intercept is the expected value of the dependent variable when all the independent variables are zero, the intercept is essentially zero in Models 2 and 3 because all the independent variables in this model are grand-mean centered, and because the dependent variable has a mean of zero. (Recall that with grand-mean centering, the independent variables all have a mean zero, and the intercept is the average value of the dependent variable). In model 3, the public and HBCU dummy variables are left uncentered, thus the intercept is the predicted level of engagement for a student with average values for the student-level variables, attending a historically White, private institution that has average values for the continuous school-level variables.

In sum, the results for Model 3 suggest that student engagement is affected by several student-level and school-level variables. The next question to be answered is, does the impact of any of the student-level variables differ across schools? Although any of the student-level variables could vary across

schools, I focus here on the dummy variable for females. As seen in Model 3, Table 4, females are on average about .13 standard deviations more engaged than males. The next set of analyses tests if this "gender gap" varies across schools, and if so, if any of this variation can be explained by the school-level variables.

The first column of Table 5 lists the results for Model 3 from Table 4, the random intercept model for student engagement. Model 5 is the same model as Model 3, with two exceptions. The female dummy variable at the student level has been group-mean centered (recall that it was grand-mean centered in Model 3), and the coefficient has been randomized. Model 6 is a random coefficient model, where a different coefficient for the effect of gender has been estimated for every college. No explanatory variables have been included for the female coefficient other than an intercept and a random term (see e.g. Equation 4.4). At the bottom of the table there is now a third variance component τ_{11}, the variance of the coefficients for the female dummy variable. The hypothesis test that this variance is equal to zero is rejected at $p<.01$, so it is possible that we can model this variance in some way.

Model 7 in the last two columns of Table 5 is a simple attempt to model the variance in the female dummy variable coefficient. Note that there are two models, each with a different dependent variable. The first column is similar to Model 6; the dependent variable is the student engagement factor, and variance in this factor is partially explained by student-level and college-level variables. The dependent variable in the second column is the female dummy variable coefficient for each school, with variation modeled with only school-level variables, as there is only one female slope coefficient per school.

One possibly confusing aspect of Model 7 is that there appears to be no coefficient for the impact of being female on engagement. Because this slope coefficient is now estimated for each school, we need the average of these coefficients to understand for the impact of gender; this impact is the intercept in the second column. Just as in the random intercept model, the intercept is the average of the school coefficients. We can see that the value of the intercept is .2797, much larger than the .16 in Model 6. Again, one must be careful in interpreting the values of intercepts and slopes in a multilevel model when they have been randomized. The value of .1594 in Model 6 is the average of each school's female-male engagement difference. The .2797 is not the average difference for all schools, but is instead the average difference in engagement between females and males in a historically White, private institution that has average values for the continuous school-level variables. If both the HBCU and public dummy variables had been grand-mean centered, the intercept in the second column would equal .16 rather than .28.

We can see from the statistical significance of the independent variables in the slope model, as well as the variance explained measure at the bottom of the table, that the model does a poor job explaining why females are more engaged than males at some schools. However, it does appear that selectivity may play a role in the impact of gender ($p < .08$).

Table 5
Correlates of Engagement, Multilevel Random Intercept and Random Coefficient Models

	Random intercept	Random coefficient model		
	3	6	7	7
Dependent variable	Engagement	Engagement	Engagement	Slope for female
Student-level (N=4,481)				
Female	.1327**	.1594**	-	-
	(4.49)	(4.22)		
Non-white	.1593**	.1706**	.1705**	-
	(4.22)	(4.53)	(4.53)	
1st generation	-.0757*	-.0825**	-.0836**	-
	(-2.51)	(-2.74)	(-2.77)	
SAT score	-.0007**	-.0007**	-.0007**	-
	(-7.50)	(-7.45)	(-7.53)	
College GPA	.0011**	.0011**	.0011**	-
	(6.11)	(6.08)	(6.11)	
Full-time	.2057**	.2052**	.2085**	-
	(4.79)	(4.75)	(4.82)	
School-level (N=360)				
Enrollment	-.0167**	-.0169**	-.0172**	-.0054
	(-2.61)	(-2.71)	(-2.76)	(-0.48)
Enrollment squared	.0002*	.0002*	.0002*	.0001
	(2.01)	(2.03)	(2.08)	(0.52)
Selectivity	.0975**	.0965**	.0990**	-.0738+
	(4.19)	(4.21)	(4.33)	(-1.79)
% female students	.1106	.2382*	.2371*	.0204
	(0.94)	(2.05)	(2.05)	'(0.07)
% graduate students	-.0920	-.0996	-.0985	.3741
	(-0.45)	(-0.50)	(-0.49)	(0.96)
HBCU	-.0895	-.1085	-.1052	-.2951
	(-0.81)	(-0.99)	(-0.97)	(-1.43)
Public	-.2714**	-.2779**	-.2707**	-.1705
	(-4.20)	(-4.39)	(-4.29)	(-1.48)
Intercept	.2041**	.2066**	.2029**	.2797**
	(4.51)	(4.65)	(4.58)	(3.41)
Variance components				
σ^2	.7915**	.7625**	.7628**	
τ_{00}	0750**	.0679**	.0667**	
τ_{11}		.1378**		.1381**
Student-level: % var. exp.	.03	.07	.07	
School-level: % var. exp.	.46	.51	.52	
Female slope: % var exp.				.00

Note: t-statistics are shown in parentheses; p<.01 **, p<.05 *. p<.10 +.
All variables are grand-mean centered except for HBCU and public; female is grand-mean centered in Model 3 and group-mean centered in Models 6 and 7.

What does the selectivity result tell us about gender differences in engagement? Because higher values of the selectivity measure indicate more selective institutions, the negative coefficient for selectivity in the slope coefficient model indicates that the female dummy variable coefficient is on average smaller at selective institutions (recall that the intercept, or average female dummy variable coefficient, is .28; for every unit increase in selectivity, this coefficient drops in value by .07). Indeed, the results indicate almost no differences in engagement between females and males at the most selective institutions, and that the gender difference increases as a school becomes less selective, thus the female-male disparity is greatest at non-competitive schools.

Conclusion

Multilevel models offer institutional researchers another statistical tool to investigate the effects of institutions upon students and faculty. While multilevel models have several advantages over OLS, two of the most important are the correct estimation of standard errors for institution-level variables, and the ability to model why the effects of individual-level variables vary across institutions. However, these advantages come at a price. Multilevel models require data from multiple groups, and such data can be difficult and costly to compile.

Besides the application reviewed in this chapter, multilevel models can be used for other analyses. The two-level model can easily be extended to three levels, such as faculty nested within departments nested within colleges, although such three-level models are currently uncommon in higher education research. The ability to deal with grouped data also allows multilevel models to be applied to other types of data. They can, for example, be used to analyze data over time, such as multiple observations of students (Hedecker, 2004; Raudenbush & Bryk, 2002, chapter 6). Instead of students nested within colleges, here observations of particular students are nested within students. From a higher education perspective, one example might be an analysis of why students' GPA varies from semester to semester, so each student would have several semesters of observations. Finally, multilevel models can also be used for meta-analyses of multiple studies (Konstantopoulos & Hedges, 2004; Raudenbush & Bryk, 2002, chapter 7).

Additional Resources

Ethington (1997) covers much of the same material in this chapter in greater detail and is an excellent short introduction to multilevel models. After reading her chapter and this one, the reader should be an educated consumer of most multilevel analyses published in higher education research journals.

The classic multilevel text is Raudenbush and Bryk (2002), and this book is a must-read for anyone wishing to begin using multilevel models for

data analysis. They cover the major applications of multilevel analysis, as well as many practical modeling issues. Heck and Thomas (2000) is another excellent multilevel text, and is particularly useful for anyone wishing to learn about multilevel structural equation models. Snijders and Bosker (1999) is a standard text often used by European scholars. For two excellent discussions of centering, see Kreft, De Leeuw and Aiken (1995) and Raudenbush and Bryk (2002, pp. 31-35 and 134-149).

Several different software packages are available to estimate multilevel models. Two of the most popular stand-alone packages are HLM (http://www.ssicentral.com) and MLwiN (http://multilevel.ioe.ac.uk/index.html). The former was written in part by Raudenbush and Bryk, and uses the same notation and terminology of their text; in addition, a student version is available for free at http://www.ssicentral.com/other/hlmstu.htm. The current list price for a single-user license is $395 for HLM and $540 for MLwiN (with the academic discount).

For many years SAS (http://www.sas.com) has had procedures available for estimating multilevel models (PROC MIXED and PROC NLMIXED). Although complicated to use, two papers clearly describe how to use SAS to estimate both linear and nonlinear multilevel models. Singer (1998) is by far the best introduction to using SAS for multilevel models, while Yang (2003) covers nonlinear models.

SPSS has recently introduced a multilevel modeling module (http://www.spss.com/advanced_models/data_analysis.htm); I have not used the module but it appears complex. Other software possibilities include a free module available for use with Stata to estimate multilevel models (http://www.gllamm.org/) as well as modules available for the free statistics package R (http://www.r-project.org/).

Besides the Web sites listed above, the following Web sites are also useful resources for multilevel modeling:

- Multilevel Modeling Newsletter - http://multilevel.ioe.ac.uk/publref/newsletters.html
- Multilevel listserv - http://www.jiscmail.ac.uk/lists/multilevel.html
- UCLA Multilevel Modeling Portal - http://statcomp.ats.ucla.edu/mlm/
- Stephen Raudenbush's homepage - http://www-personal.umich.edu/~rauden/
- Tom Snijders' homepage - http://stat.gamma.rug.nl/snijders/multilevel.htm

References

Ethington, C. A. (1997). A hierarchical linear modeling approach to studying college effects. In J. C. Smart (Ed.), *Higher Education: Handbook of Theory and Research* (Vol. 12): Agathon Press.

Heck, R. H., & Thomas, S. L. (2000). *An Introduction to Multilevel Modeling Techniques*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hedecker, D. (2004). An introduction to growth modeling. In D. Kaplan (Ed.), *Sage Handbook of Quantitative Methodology for the Social Sciences*. Thousand Oaks, CA: Sage Publications.

Hu, S., & Kuh, G. D. (2002). Being (dis)engaged in educationally purposeful activities: The influences of student and institutional characteristics. *Research in Higher Education, 43*(5), 555-575.

Hu, S., & Kuh, G. D. (2003a). Diversity experiences and college student learning and personal development. *Journal of College Student Development, 44*(3), 320-334.

Hu, S., & Kuh, G. D. (2003b). Maximizing what students get out of college: Testing a learning productivity model. *Journal of College Student Development, 44*(2), 185-203.

Johnsrud, L. K., & Rosser, V. J. (2002). Faculty members' morale and their intention to leave. *Journal of Higher Education, 73*(4), 518-542.

Kennedy, P. (2003). *A Guide to Econometrics*. Cambridge, MA: MIT Press.

Kim, M. M. (2002a). Cultivating intellectual development: Comparing women-only colleges and coeducational colleges for educational effectiveness. *Research in Higher Education, 43*(4), 447-481.

Kim, M. M. (2002b). Historically Black vs. White institutions: Academic development among Black students. *Review of Higher Education, 25*(4), 385-407.

Konstantopoulos, S., & Hedges, L. V. (2004). Meta-analysis. In D. Kaplan (Ed.), *Sage Handbook of Quantitative Methodology for the Social Sciences*. Thousand Oaks, CA: Sage Publications.

Kreft, I. G. G., De Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research, 30*(1), 1-21.

Kuh, G. D., & Hu, S. (2001). The effects of student-faculty interaction in the 1990s. *Review of Higher Education, 24*(3), 309-332.

Merlo, J., Ostergren, P.-O., Hagberg, O., Lindstrom, M., Lindgren, A., Melander, A., et al. (2001). Diastolic blood pressure and area of residence: Multilevel versus ecological analysis of social inequity. *Journal of Epidemiology and Community Health, 55*(11), 791-798.

Porter, S. R., & Umbach, P. D. (2001). Analyzing faculty workload data using multilevel modeling. *Research in Higher Education, 42*(2), 171-196.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Liner Models: Applications and Data Analysis Methods*. Thousand Oaks, CA: Sage Publications.

Rosser, V. J., Johnsrud, L. K., & Heck, R. H. (2003). Academic deans and directors: Assessing their effectiveness from individual and institutional perspectives. *Journal of Higher Education, 74*(1), 1-25.

Rumberger, R. W., & Thomas, S. L. (1993). The economic returns to college major, quality and performance: A multilevel analysis of recent graduates. *Economics of Education Review, 12*(1), 1-19.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 24*(4), 323-355.

Smyth, F. L., & McArdle, J. J. (2004). Ethnic and gender differences in science graduation at selective colleges with implications for admission policy and college choice. *Research in Higher Education, 45*(4), 353-381.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Thousand Oaks, CA: Sage Publications.

Strauss, L. C., & Volkwein, J. F. (2004). Predictors of student committment at two-year and four-year institutions. *Journal of Higher Education, 75*(2), 203-227.

Thomas, S. L. (2000). Deferred costs and economic returns to college major, quality, and performance. *Research in Higher Education, 41*(3), 281-313.

Toutkoushian, R. K., & Smart, J. C. (2001). Do institutional characteristics affect student gains from college? *Review of Higher Education, 25*(1), 39-61.

Umbach, P. D., & Kuh, G. D. (in press). Student experiences with diversity at liberal arts colleges: Another claim for distinctiveness. *Journal of Higher Education*.

Umbach, P. D., & Porter, S. R. (2002). How do academic departments impact student satisfaction? Understanding the contextual effects of departments. *Research in Higher Education, 43*(2), 209-234.

Yang, M. (2003). *A Review of Random Effects Modeling in SAS (Release 8.2)*. London: Centre for Multilevel Modeling.

Endnotes

[1] Raudenbush and Bryk (2002), pp. 117-118 discuss the technical issues of using this approach with OLS.

[2] Note that some scholars refer to level 1 as the micro-level, and level 2 as the macro-level (see Snijders & Bosker, 1999, p. 8).

Chapter 5
Identifying and Analyzing Group Differences

Victor M. H. Borden

Many of the issues considered by institutional researchers involve the classification of current and potential students, faculty, academic programs, or higher education institutions into groups. Examples include identifying potential student markets, predicting which students will most likely persist in their studies, determining which students are most likely to benefit from various support programs, and selecting peer institutions. Discriminant analysis, logistic regression, and cluster analysis are among the most suitable techniques for addressing these classification issues.

This chapter considers two sides of the "group differences and classification" coin: analyzing differences among existing groups (e.g., students accepted for admission who choose to enroll at one's institution versus those who choose to enroll elsewhere, retained versus non-retained first-year students, etc.); and identifying groups within previously undifferentiated populations (e.g., peer institutions, student market segments, etc.).

Discriminant analysis and logistic regression consider the issue of how a set of predictors can best distinguish among members of pre-existing groups. For these methods, the dependent variable is a nominal categorization of objects (group membership) that is known, a priori. The predictors can be continuous or categorical (nominal) variables. Between the two techniques, logistic regression is currently more popular because of its robustness and superiority in handling categorical predictors. However, discriminant analysis is better suited for grouping outcomes that have more than two unordered values as, for example, in a retention analysis, if one wanted to distinguish between students who returned for their sophomore year, students who transferred to another institution, and students who are not attending college.

Cluster analysis refers to a broad variety of techniques for determining if the objects of study (people, institutions, programs, etc.) can be grouped in a meaningful way such that members of a group are relatively similar to each other and relatively different from members of other groups. That is, clustering techniques are used to define groups that do not exist, a priori. Cluster analysis techniques are not based on parametric statistics, but rather on numerical heuristics. As a result, they are less consistent and more sample-specific than most parametric statistical techniques. They are also computationally intensive and so have not been practical until relatively recently. Some cluster analysis algorithms will tax even modern-day computers.

Grouping and classification analyses can be quite useful in institutional research. Enrollment management has been the most fertile area for this type of analysis, including identifying niche markets for specific types of academic programs or support program development, analyzing admitted students who enroll versus those who do not, predicting returning versus non-returning students, graduating versus non-graduating students, those likely to donate versus not donate and so on. These techniques can also be applied to faculty salary and workload studies, staff vacancy issues, corporate donations, and any other issue where the objective is to classify people or other objects (e.g., programs, classes, facilities) for planning and evaluation purposes.

Analyzing Differences among Existing Groups

The t-test is the first statistic one usually learns related to group differences. For this statistic, group membership defines the independent variable, which is then tested for its impact on a single continuous dependent variable. For example, one may want to know whether studying or not studying for an exam (the independent variable or treatment condition) results in significantly different grades (the dependent variable or outcome). One initially learns that a t-test is valid only in an experimental situation, where assignment to the treatment condition is random. However, one soon learns that the t-test is "robust" enough under violations of basic assumptions that it can be used to test for differences among existing groups, such as between men and women, or undergraduate and graduate students.

Does this mean that one can use a t-test to analyze differences between retained and non-retained students across such factors as their college entry exam scores or their average grades during their first year of college? Conducting such tests could provide some useful information regarding how retained and non-retained students differ. However, there is a fundamental difference between examining group differences where the group characteristic is considered the independent variable and where it is considered the dependent variable. In the t-test design, group membership is the independent variable and is seen as "causing" the differences in the dependent variables. Obviously, this is not an appropriate characterization of the retention issue.

When thinking about analyzing group membership as an outcome or dependent variable, one might first think about using linear regression, where a set of predictors can be tested for their impact on a single outcome variable, such as whether a student is retained or not retained. However, simple linear (ordinary least squares, or OLS) regression requires the dependent variable to have some characteristics that a group outcome violates. First, the dependent variable in a linear OLS regression must have an unlimited range but a group classification can take on only two discrete values, which we usually designate as one or zero[1]. If one uses linear OLS regression to

predict such a dichotomous outcome, one could get predicted values greater than one or less then zero, which are impossible. Second, the coefficients in a linear OLS regression are additive, meaning that it can be observed how the impact of each predictor "adds to" the prediction of differences in the outcome. However, if the effects expressed in coefficients to predict a group outcome are added together, the resulting values can range continuously from well below zero to well beyond one, making it difficult to interpret how the predictors contribute to the outcome. Having a dichotomous outcome violates several other important assumptions of linear OLS regression, including that the variance of the outcome is constant across values of the predictors (homoscedasticity), and that the differences between the predicted and actual values of the outcome (i.e., the error term) are normally distributed.

Discriminant analysis and logistic regression are two regression techniques that address these issues and thus allow for group membership as a dependent variable. Before examining each specific method in some detail, we first consider the common aims of these techniques and how they apply to some typical institutional research issues.

The Common Aims of Discriminant Analysis and Logistic Regression

These two techniques address the same three basic questions:

1. To what extent do specific predictors contribute to determining the group membership outcome?
2. What is the best combination of predictors to optimize predictions of the group outcome?
3. How useful is that combination for classifying new cases?

The first two of these questions are the traditional linear OLS regression questions, focusing on the contribution of each predictor (the beta coefficients) and the strength of the overall prediction model (the F-ratio for the entire equation and the model R-Square value), respectively. The third issue of classification corresponds to the prediction capabilities of a linear OLS regression model. However, the method by which prediction is characterized differs substantially. Within linear OLS prediction, the differences between predicted and actual values are characterized as a deviation score, which then can be summarized through an overall error term. For discriminant analysis and logistic regression, prediction can also be characterized according to the percentage of cases that are correctly or incorrectly classified.

Both discriminant analysis and logistic regression use an outcome transformation technique to get around the "linear OLS problem" when using group outcomes. However, each of these techniques uses a different transformation approach. The discriminant analysis workaround is to predict a "discriminant function" that is, a linear combination of the predictors that generates the largest mean differences between the groups. This outcome,

typically characterized as "D," is a continuous variable for which the average for members of one group is as far from the average for members of the other group as the given set of predictors will allow. In logistic regression, the transformed outcome used in the statistical technique is the natural log of the quantity: the probability of belonging to one group divided by the probability of belonging to the other group. This is called the log of the odds ratio, or the logit.

By using these outcome transformation techniques, discriminant analysis and logistic regression provide a statistically valid way to assess the impact of a set of predictors on a group membership outcome. However, in doing so the predictor coefficients they yield and the overall model fit statistics are not as easy to interpret as are linear OLS regression coefficients and model statistics.

Another common feature of these two techniques is a requirement that the grouping characteristic must be mutually exclusive (i.e., one group cannot be a subset of another). In addition, and in common with linear OLS regression, these techniques require that the predictors not be linearly dependent (i.e., math SAT score, verbal SAT score and Total SAT score can not be used as predictors in the same model because Total SAT score is a linear combination of math and verbal).

As with other regression techniques, both discriminant analysis and logistic regression allow the predictors to be entered into the model all at the same time or in one of two stepwise fashions: forward (starting with the strongest predictor, adding the next strongest and so on) or backward (starting with all predictors, removing the weakest predictor, removing the second weakest predictor and so on). Several different criteria are available for these stepwise methods to determine which variables to enter or remove, and to determine when to stop entering or removing variables. Version 12.0 of the SPSS® statistical software package, which will be used in the examples reviewed in this chapter, includes simultaneous predictor entry for both logistic regression and discriminant analysis. For stepwise entry, SPSS® 12.0 offers three different criteria each for forward and backward entry of predictors into a logistic regression and five different criteria, but for only forward entry, for discriminant analysis.

Typical Institutional Research Questions Addressed by These Techniques

The discussion to this point referred several times to admissions yield and retention as two common institutional research questions that can be explored with these techniques. Also mentioned in the introduction is an issue in which university development staff would be quite interested: what predicts alumni donation behavior (i.e., which alumni will and will not donate).

Another common institutional research issue that can be addressed through these techniques is participation in academic support or student life programs. For example, these methods can be used to determine the

differences between students who take advantage of supplemental instruction programs versus those who do not. Or, one can use these techniques to determine which students are most likely to study abroad, participate in an internship program, or seek help from a career counseling center.

Participation questions are interesting in and of themselves, and also as the "self-selection" component in a multi-stage analysis of intervention impact. That is, if one can predict participation in a program, then one can control for self-selection when assessing the impact of that program, by using the prediction of participation as one of the predictors in the impact assessment[2].

In addition to student behavior, these techniques can be used to explore differences in faculty behavior (e.g., retention, obtaining tenure, receiving research grants, publishing, serving on university committees, etc.) and staff behavior (retention, benefit package choices, participating in training and development, etc.).

Discriminant analysis and logistic regression can also be used to explore differences between groups of classes, departments, facilities, financial aid packages, institutions, states, and so on. Table 1 provides examples of the variety of research questions that can be addressed for various types of entities within higher education.

Discriminant Analysis

Discriminant analysis follows quite closely the linear OLS regression framework. The technique determines the linear combination of predictor variables that accounts for the most variation in the group membership outcome. As with linear OLS regression, the results provide statistics related to the overall fit of the model and as to which of the individual variables contribute significantly to the overall model.

The basic equation for discriminant analysis, called the canonical discriminant function, is shown in equation (5.1). D is the discriminant function score for a given case (person or object), B_i is the discriminant coefficient for the i^{th} predictor variable, and X_i is the value of the i^{th} predictor variable for that case.

$$D = B_0 + B_1X_1 + B_2X_2 + \ldots + B_pX_p \qquad [5.1]$$

Discriminant analysis requires two or more groups, at least two cases (observations) per group, and a maximum of n-2 predictor variables, although the power of the technique is greater if there are many more cases than predictor variables. Technically, the predictor variables should be on an interval or ratio scale. However, just like with linear OLS regression, the technique is robust enough to handle dichotomous (dummy) predictors and, some would argue, scale variables comprised of multiple ordinal variables (e.g., a satisfaction measure that is derived from the sum of Likert-scale items, as

Table 1
Examples of Research Questions about a Variety of Entities That Can be Addressed by Discriminant Analysis and Logistic Regression

Entity	Research Questions
Students	What factors determine whether students enroll, persist, participate in support programs, engage in enriched learning opportunities, etc?
Faculty	What factors are associated with faculty being retained, attaining tenure, obtaining research grants, teaching online courses, etc?
Staff	What determines whether staff members are retained, seek training opportunities, choose specific benefit options, etc?
Classes	What factors determine whether a class is sufficiently subscribed; scheduled in the day or evening; face-to-face or online; on-campus or off-campus?
Academic Departments	What predicts whether departments adopt assessment programs? How do student profiles differ between specific departments?
Facilities	How do older and newer facilities differ in their costs of operation?
Financial Aid Packages	What kinds of students obtain primarily merit versus need aid?
Budgets	How do expenditure patterns differ in academic versus administrative unit budgets?
Feeder High Schools	How does the student profile differ for students who originate from high schools in the immediate region versus those from other parts of the state or country?
Higher Education Institutions	How do public and private institutions differ in the diversity of their students (race, age, course load, etc.)
States	How do high tuition/high aid states differ from low tuition/low aid states in terms of student access and success?

long as the scale has high internal reliability). As mentioned earlier, no predictor variable can be a linear combination of other predictor variables.

Discriminant analysis also shares with linear OLS regression the assumption of homogeneity of variance, which requires that the variance-covariance matrices for the different groups must be approximately equal. However, several adjustment formulas have been developed to accommodate situations where this assumption is violated. Finally, as with linear OLS regression, the observations must be drawn from a population with a multivariate normal distribution on the predictor variables.

Table 2
Predictor Variables Considered in the Retention Example Used for Discriminant and Logistic Regression Analyses

Code Name	Description	Values
SEMGPA	Fall Semester Grade-Point Average	Ranges from 0 to 4; mean=2.57, SD=1.09; median=2.85
CRSLOAD	Fall Semester Credit Load	Ranges from 1 to 18; mean=12.3, SD=3.1; median=13
FEMALE	Gender	Dummy variable where 0=male (44%) and 1=female (56%)
AGE	Student age at entry	Ranges from 16 to 56; mean=20.1, SD=5.4, Median=18
SATACT	Total SAT score, or ACT total converted to SAT scale	Ranges from 480 to 1470; mean=991, SD=147; median=990
HSPCT	Percentile rank in high school	Ranges from 0 (lowest rank) to 99.7 (highest rank); mean=59.9, SD=22.1, Median=61
CLPREP	Number of college preparatory units completed in high school	Ranges from 3 to 51; mean=33.8, SD=6.1, Median=34
CNDADM	Whether the student was admitted conditionally or unconditionally	Dummy variable where 0=admitted unconditionally (48%) and 1=admitted conditionally (52%)

The various components of output produced by a discriminant analysis are reviewed next using a concrete example: predicting retention to the second year of first-time, first-year students. Table 2 summarizes the predictors considered in this analysis.

The specific output derived from a discriminant analysis depends on the various options that are chosen. For example, one can request univariate descriptive statistics on the predictors and univariate ANOVAs related to how

each predictor relates to the group outcome, as well as the variance-covariance matrix among the predictors for all subjects or for each group separately. None of these are included in the current example.

Table 3 shows the case processing summary from the discriminant analysis, showing that 1,538 of the 2,419 observations in the data set had non-missing values for all variables included in the analysis. Table 4 shows the results for Box's test of equality of covariance matrices, which should be

Table 3
Discriminant Analysis Case Processing Summary

Unweighted Cases		N	Percent
Valid		1538	63.6
Excluded	Missing or out-of-range group codes	0	0
	At least one missing discriminating variable	881	36.4
	Both missing or out-of-range group codes and at least one missing discriminating variable	0	.0
	Total	881	36.4
Total		2419	100.0

run to determine if the homogeneity of covariance assumption has been met. In this case, the very small p-level indicates that this assumption has not been met, which should be noted as a caveat when interpreting the remaining output.

The first substantive results for the discriminant analysis are two small tables that provide information relating to the overall performance of the tested model. Table 5 includes both the eigenvalue/canonical correlation table and the Wilk's Lambda table. The current example uses as an outcome the dichotomous (two-group) distinction: retained or not retained. Discriminant analysis can also be used to test

Table 4
Results from Box's Test of Equality of Covariance for the Discriminant Analysis

retn	Rank	Log Determinant
0	8	17.774
1	8	17.149
Pooled within-groups	8	17.672

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

Box's M		482.229
F	Approx.	13.310
	df1	36
	df2	3786143.240
	Sig.	.000

Tests null hypothesis of equal population covariance matrices.

polychotomous (multi-group) outcomes. However, when there are more than two groups, there is more than one resulting function (i.e. more than one solution). Generally, if the outcome has *k* groups, the analysis produces *k*-1 functions, where each function is orthogonal (uncorrelated) to the others. Because the current example has a dichotomous outcome, Table 4 shows a single function that accounts for 100% of the model's total variation (as reflected in the third and fourth columns of the top part of Table 5).

Table 5
Overall Model Statistics for the Discriminant Analysis

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.309(a)	100.0	100.0	.486

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.764	412.553	8	.000

The only informative information left in the top portion of Table 5 is the eigenvalue and canonical correlation of the single discriminant function produced by the model. The eigenvalue is the ratio of the sum of squares between groups divided by the sum of squares within groups (i.e., analogous to the treatment effect in a one factor ANOVA). The final parameters of a discriminant function are those that maximize the eigenvalue. The canonical correlation is a version of the standard Pearson product-moment correlation coefficient, where the values being correlated are subjects' actual group membership and the predicted D-values (i.e., multiplying each variable by its coefficient and summing across variables). Since this is a correlation between a dichotomous and continuous variable, it is technically a point-biserial correlation, which is restricted in range to values slightly below 1.00 (depending on the sample size and distribution of data). With this limitation in mind, the canonical correlation for a discriminant function is much like the multiple correlation (R) in a linear OLS regression, except it is slightly attenuated. Also, like the multiple correlation, its squared value can be taken as a measure of the overall fit of the model (i.e., the percent of total variation in the outcome that is accounted for by the predictors).

But discriminant analysis does not typically use the squared canonical correlation to express the overall model fit. Rather, it employs Wilk's Lambda—a common multivariate linear model statistic that expresses the unaccounted for variation in the model. Wilk's Lambda is used to test the null hypothesis that the populations from which the samples came had identical average *D* values. Thus, if Wilk's Lambda is 1.0, then the model suggests

that there is no difference between groups on the composite predictors. In this example, Wilk's Lambda is 0.726, which is significantly different from 1.0 and shows that the model does not account for 72.6% of the variation in the group outcome. The complement to Wilk's Lambda (1 - .726, or .274 in this case), is the equivalent to the R^2 of linear OLS regression and is also equal to the square of the canonical correlation.

The next three components of the discriminant analysis routine are here combined into the columns of Table 6. The first column of numbers represents the standardized predictor coefficients. That is, these coefficients would be applied to the values of the predictor variables that have been transformed to z-scores to yield the overall discriminant function value. The standardized coefficients, like the beta weights of a linear OLS regression, show the relative contribution of each variable to the overall prediction model, taking into account differences in scale. This is useful for model building, but less useful for prediction. The unstandardized coefficients provide more practical information as they express how the overall discriminant function changes for each unit change in the raw (untransformed) predictor variable. In this example, a one point grade increase (1.00) in semester GPA would

Table 6
Discriminant Analysis Predictor Coefficients (Standardized and Unstandardized) and Structure Matrix

	Standardized Coefficients	Unstandardized Coefficients
SEMGPA	.948	1.042
CRSLOAD	.260	.116
FEMALE	.099	.200
AGE	.069	.047
SATACT	-.031	.000
HSPCT	.027	.001
CLPREP	.013	.003
CNDADM	-.054	-.109
(Constant)		-5.062

Structure Matrix	
SEMGPA	.958
HSPCT	.271
CNDADM	-.249
CRSLOAD	.219
FEMALE	.149
SATACT	.134
CLPREP	.101
AGE	.028

result in an increase in 1.042 of the overall discriminant function value. Unfortunately, the overall unstandardized discriminant function value does not have direct meaning and thus unstandardized coefficients are often ignored in discriminant analysis.

The final two columns of Table 6 show the structure matrix for the discriminant function. The structure matrix is comprised of the correlations (Pearson product-moment) between each predictor and the overall discriminant function value (*D*). The listing of the predictor variables in the structure matrix is sorted so that the predictor with the highest correlation (absolute value) is at the top of the list and the predictor with the lowest correlation is at the bottom of the list. For the standardized and unstandardized coefficient tables, the predictors are ordered as indicated when entered into the model.

In addition to the overall and coefficient statistics, discriminant analysis models can be judged according to their ability to accurately classify observations. This can be done on the sample from which the model was estimated or, even better, against a second sample of data. If one has enough cases, it is usually recommended that the sample be split so that the model can be developed using one portion of the sample and then tested for accuracy using the second portion of the sample.

Table 7 shows the results of classification for the current model on both the cases upon which the model was built, as well as the "Cases Not Selected," which were set aside for validation purposes. The results reveal that the model correctly classified 74.2% of the cases included in the analysis (80% of retained students and 62% of non-retained students classified correctly). The classification accuracy was similar for the non-selected cases: 72.8%

Table 7
Classification Results from Discriminant Analysis

				Predicted Group Membership		
			retn	0	1	Total
Cases Selected	Original	Count	0	154	95	249
			1	108	431	539
		%	0	61.8	38.2	100.0
			1	20.0	80.0	100.0
Cases Not Selected	Original	Count	0	174	91	265
			1	113	372	485
		%	0	65.7	34.3	100.0
			1	23.3	76.7	100.0

a 74.2% of selected original grouped cases correctly classified.
b 72.8% of unselected original grouped cases correctly classified.

overall including 77% accuracy among retained students and 66% accuracy among non-retained students. Thus the current model, although relatively weak in its predictive capacity, is reliable (i.e., consistent) across samples.

Discriminant Analysis References

The following references are recommended for exploring further details on the method of discriminant analysis and its application to institutional research:

Institutional Research Applications

Krotseng, M. V. (1992). Predicting persistence from the student adaptation to college questionnaire: Early warning or siren song? *Research in Higher Education, 33*, 99-111.

Riggs, M., Downey, R,. McIntyre, P., & Hoyt, D. (1986). Using discriminant analysis to predict faculty rank. *Research in Higher Education, 25*, 365-376.

Urban, R. F. (1992). Increasing admitted student yield using a political targeting model and discriminant analysis: An Institutional Research-Admissions partnership. *AIR Professional File, No. 45.* Tallahassee, FL:

General References

Dillon, W. and Goldstein, M. (1984), *Multivariate analysis: Methods and applications*, New York: Wiley

Hand, D. J. (1981). *Discrimination and classification*, New York: Wiley.

Huberty, C.J. (1984). Issues in the use and interpretation of discriminant analysis. *Psychological Bulletin, 95*, 156-171.

Klecka, W. R. (1980). Discriminant analysis. *Quantitative Applications in the Social Sciences, No 19.* Beverly Hills, CA: Sage.

Lachenbruch, P. A. (1975). *Discriminant analysis.* New York: Hafner Press.

Logistic Regression

Like discriminant analysis, logistic regression uses the same general framework as linear OLS regression, wherein coefficients are estimated to provide the linear combination of predictors that best predicts the group outcome. Logistic regression departs from linear OLS regression in two fundamental ways. First, rather than using the minimization of the error term as the criteria for determining the best combination of predictors (i.e., minimizing the squared distance between the observed and predicted values), logistic regression uses the maximum likelihood (MLE) approach to estimation, where coefficients are chosen that produce the greatest probability of obtaining that particular set of data given the fitted regression coefficients.

But the more important difference between linear OLS regression and logistic regression is the outcome transformation that is employed to get

around the problematic distribution properties of a dichotomous outcome. As mentioned earlier, logistic regression transforms the outcome into the log of the odds ratio or logit, which is the natural log of the quantity: the probability of belonging to one group divided by the probability of belonging to the other group. For example, if *P* represents the probability of being retained from the first to the second year of college (i.e., the first-year retention rate), and 1-*P* represents the attrition rate (probability of not being retained), then the logit is represented by equation 5.2

$$\text{logit}(P) = \log(\text{odds ratio}) = \ln\left(\frac{P}{1-P}\right) \qquad [5.2]$$

The logistic regression equation relates this value to the predictors in the same way that the OLS regression relates a raw outcome value to a set of predictors, as shown in equation 5.3

$$\text{logit}(P) \quad a \quad b_1X_1 \quad b_2X_2 \quad \eth K \quad b_pX_p \qquad [5.3]$$

where a is the y-intercept and the b_1 through b_p are the regression coefficients. Just as in linear OLS regression, the b coefficients represent the unit change in the outcome (in this case logit(*P*)) for each unit change in the predictor.

Because it is difficult to directly understand unit changes in the log of the odds ratio, we usually relate our results to the exponent or anti-natural log of the coefficient (i.e., Exp(b) or e^b). This is more interpretable, since it reflects how a unit change in the predictor influences the odds ratio rather than the log of the odds ratio. Put more simply, the exponent of the coefficient for a given predictor reveals how many times more likely it is to obtain the outcome (e.g., to be retained) for each unit change in the predictor. For example, consider predicting student retention (one for retained, zero for not retained) from gender (where one represents female and zero male) and obtaining a b coefficient of 0.223. The exponent of this coefficient, exp(b)=1.25 tells you that female students are 1.25 times more likely to be retained than males.

The coefficients derived from a logistic regression, like their linear OLS counterparts, are error-laden estimates of the "true" population coefficient. As such, they are subject to significance tests to determine if they are statistically significant, that is, different enough from zero to be considered a real (population) effect. The Wald statistic is used in logistic regression to express the significance level of each coefficient. The Wald statistic applies equally to the odds ratio (the exponent of b) as it does to the coefficient itself (the log of the odds ratio).[3]

We have focused so far on the estimates of the coefficients. With all regression models, however, we first consider the overall significance of the model. With the linear OLS model, we look at the overall F-ratio associated with the model and then the R^2 to determine the strength of the overall model

(percentage of total variance accounted for in the outcome). With the logistic regression equation, the overall significance is determined through a model chi-square statistic derived from the likelihood of observing the obtained data given that the model specified is accurate. Specifically, the model chi-square is related to negative two times the log of this likelihood, sometimes referred to as -2*LL*. The chi-square compares -2*LL* for the model against -2*LL* for the null hypothesis model (where the coefficients have no effect). The fact that the ratio of -2*LL* for various models has a chi-square distribution allows us to compare different models, which is very useful for stepwise analyses.

One can also derive measures that are similar to the R^2 of OLS regression. For example, SPSS provides two such measures, the Cox & Snell R^2 and the Nagelkerke R^2. There are several other such Pseudo-R^2 measures, which can vary markedly for the same model. Because of their variability, and the fact that none parallel precisely the R^2 of linear OLS regression, they should be interpreted with great caution.

Tables 8 through 12 show the results of a logistic regression for our retention example. The first three tables (8, 9, and 10) provide information prior to the testing of the logistic regression model. The case processing summary of Table 8 indicates the number of cases that have valid values for all variables (outcome and predictors) versus those that have at least one missing value. Table 9 represents the classification table prior to the analysis, which shows the actual number and percentage of cases that fall into the two outcome categories, but does not yet have the predicted outcomes. Table 10 lists the predictor variables that will be considered in the model and indicates which among them are significantly associated

Table 8
Logistic Regression Case Processing Summary

Unweighted Cases(a)		N	Percent
Selected Cases	Included in Analysis	1538	63.6
	Missing Cases	881	36.4
	Total	2419	100.0
Unselected Cases		0	.0
Total		2419	100.0

Table 9
Pre-Analysis Logistic Regression Classification Table

Observed			Predicted		
			retn		Percentage Correct
			0	1	
Step 0	Retn	0	0	514	.0
		1	0	1024	100.0
	Overall Percentage				66.6

a Constant is included in the model.
b The cut value is .500

Table 10
Pre-Analysis Logistic Regression Model and Variable Statistics

Variables not in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.689	.054	162.577	1	.000	1.992

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	semgpa	339.932	1	.000
		crsload	22.416	1	.000
		female	10.542	1	.001
		age	.369	1	.543
		satact	8.437	1	.004
		hspct	34.212	1	.000
		clprep	4.841	1	.028
		cond	28.830	1	.000
	Overall Statistics		363.089	8	.000

with the outcome, not yet taking into account the other predictors. Note that in our case, age is the only prospective predictor that is not related to the outcome prior to developing the model.

The overall model statistics and logit coefficient statistics are presented in Table 11. SPSS provides the -2*LL* value as well as the two pseudo-R^2 statistics described above. It is interesting to note that the R^2 obtained through the Discriminant Analysis (1-Wilk's Lambda) or .236 is between the two values provided in the top portion of Table 11, although closer to the Cox & Snell value of .219.

The coefficient information provided in the bottom portion of Table 11 shows results very similar to our discriminant analysis on these same data. Semester GPA and Course Load are the strongest predictors of retention, followed by gender, age, and the three academic background indicators. But unlike the discriminant analysis, the logistic regression shows us that only the semester GPA and course load coefficients are statistically different from zero (at the .05 level). The exponent of the semester GPA predictor suggests that a one point (i.e., a full grade) difference in GPA results in a student being nearly three times more likely to be retained.

The last component of the logistic regression output, the final classification table shown in Table 12, shows how the model accurately classifies just under one half of the students who were not retained and just

Table 11
Logistic Regression Model and Variable Results

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1579.583(a)	.219	.304
a Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.			

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	semgpa	1.068	-.070	233.412	1	.000	2.909
	crsload	.132	.028	22.453	1	.000	1.142
	female	.245	.132	3.458	1	.063	1.277
	age	.070	.053	1.769	1	.183	1.073
	satact	.000	.000	.038	1	.846	1.000
	hspct	.002	.004	.379	1	.538	1.002
	clprep	.003	.014	.061	1	.805	1.003
	cond	-.125	.154	.659	1	.417	.882
	Constant	-5.168	1.355	14.547	1	.000	.006
a Variable(s) entered on step 1: semgpa, crsload, female, age, satact, hspct, clprep, cond.							

Table 12
Logistic Regression Final Classification Table

Observed			Predicted		
			retn		Percentage Correct
			0	1	
Step 1	Retn	0	245	269	47.7
		1	86	938	91.6
	Overall Percentage				76.9
a The cut value is .500					

over 90 percent of the students who were retained. Compared to the classification made by the discriminant model (shown in Table 7), the logistic model was considerably more accurate for the higher probability event (being retained), but considerably less accurate for the low probability event.

Logistic Regression References

Institutional Research Applications

Berge, D. A., & Hendel, D. D. (2003). Using logistic regression to guide enrollment management at a public regional university. *AIR Professional File, No. 86.* Tallahassee, FL: Association for Institutional Research.

Cabrera, A. F. (1994). Logistic regression analysis in higher education:

An applied perspective. In J. C. Smart (ed.), *Higher education: Handbook of theory and research, Vol. 10.* New York: Agathon, 225-256.

Colbeck, C. (2002). Assessing institutionalization of curricular and pedagogical reforms. *Research in Higher Education, 43*, 397-421.

DesJardins, S. L. (2001). A comment on interpreting odds-ratios when logistic regression coefficients are negative. *AIR Professional File, No. 81.* Tallahassee, FL: Association for Institutional Research.

Peng, C. J., So, T. H.. Stage, F. K., & St. John, E. P. (2002). The use and interpretation of logistic regression in higher education journals: 1988–1999. *Research in Higher Education, 43*, 259-293.

General References

Aldrich, J. H., & Nelson, F. D. (1984). *Linear probability. logit, and probit models.* Beverly Hills, CA: Sage.

Fox, J. (2000). Multiple and generalized nonparametric regression. *Quantitative Applications in the Social Sciences Series, No.131.* Beverly Hills, CA: Sage.

Fox, J. D. (1984). *Linear statistical models and related methods.* New York: Wiley.

Hosmer, D. W. (2000). *Applied logistic regression.* New York: Wiley.

Kleinbaum, D. G. (2002). *Logistic regression: A self-learning text.* New York: Springer.

Menard, S. (2001). Applied logistic regression analysis (2nd Ed.). *Quantitative Applications in the Social Sciences, No 106.* Beverly Hills, CA: Sage.

Peng, C. J.,; Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *Journal of Educational Research, 96*, 3-14.

Press, S.J. and Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, 73, 699 -705.

Choosing between Discriminant Analysis and Logistic Regression

As mentioned above, logistic regression is now a more popular choice than discriminant analysis for analyzing the impact of predictors on a dichotomous outcome. Logistic regression is generally a more flexible method that better accommodates a broader range of predictors. In both cases, the coefficients are not as easy to interpret as in linear OLS regression, but the classification tables provide a fairly intuitive mechanism for demonstrating the accuracy of the resulting models.

It was also mentioned that discriminant analysis can handle polychotomous outcomes (i.e., more than two groups) while logistic regression cannot. However, logistic regression is part of a family of analyses that can

be used for polychotomous outcomes (multinomial logistic regression) as well as rank order outcomes (ordinal logistic regression). If one is dealing exclusively with continuous predictors, either method works equally well and both may be used to examine the consistency of results. More generally, though, logistic regression is preferred.

General References

Shott, S. (1991). Logistic regression and discriminant analysis. *Journal of the American Veterinary Medical Association, 198*, 1902-1905.

Wright, R.E. (1995). Logistic regression, in L.G. Grimm & P.R. Yarnold (Eds.), *Reading and understanding multivariate statistics*. Washington, DC: American Psychological Association.

Identifying Groups within Previously Undifferentiated Populations

Cluster analysis refers to any of a wide variety of numerical procedures used to create a classification scheme, that is, to place objects into identifiable groups. Conceptually, cluster analysis is relatively easy to understand and well suited to a variety of segmentation activities. However, it involves the use of heuristic algorithms that are not generally supported by extensive statistical reasoning. In effect, cluster analysis is an entirely data driven exercise which can yield inconsistent results across samples and cannot be used to infer to populations with any degree of known certainty. Cluster analysis is considered to be among the class of techniques now referred to as "data mining."

Intuitively, clusters can be thought of as a set of objects or points that are relatively close to each other and relatively far from points in other clusters. One of the most common applications of cluster analysis in higher education is the determination of peer institutions. That is, can colleges and universities be placed into groups such that members of a specific group are relatively similar to each other and relatively different from members of other groups? The Carnegie Classification system can be viewed as a set of institutional clusters, which begs the questions: by what criteria are the institutions considered similar to or different from each other? In the case of the Carnegie Classification (at least the system in effect prior to the 2005 revision), institutions are first distinguished according to their highest degree level. Within each degree level, different characteristics are used to further distinguish institutions.

Traditional cluster analysis methods do not allow one to use different characteristics at different levels of the analysis, although the decision tree method reviewed in the final section of this chapter does accommodate such distinctions.

Traditional cluster analysis requires the simultaneous use of a single set of variables for grouping all observations. These variables are used to construct some measure of similarity or distance, which is then processed through one of several possible clustering algorithms.

Selecting Variables

The most popular forms of cluster analysis are based on measures of "similarity" among objects according to some combination of attributes. For example, higher education institutions are typically classified according to their size, the level and types of degrees conferred, student enrollment characteristics (such as percent of full-time students, or of women or minority students), and financial characteristics (such as expenditures for instruction or research activities). One could also consider classifying students at an institution according to such characteristics as personal or family demographics, levels of academic preparation, attitudes and interests, expectations and goals, program of study, college performance, etc.

The choice of variables is one of the most critical steps in the cluster analysis process that should be guided by an explicit theory or at least solid reasoning. For example, the specification of peer institutions will vary widely depending on what characteristics are considered. It is imperative that the researcher sufficiently describes the rationale for selecting input characteristics, as well as the limitations inherent in any specific set.

Choosing a Distance/Similarity Measure

The variables selected for the analysis are used to derive a single measure of similarity among all of the cases (e.g., among each pair of students or each pair of institutions). The derived measure can be either one of distance (geometric distance between points in a multi-dimensional space) or similarity (association or correlation coefficients).

The type of variables chosen for analysis constrains the choice of similarity measure. For continuous variables, such as age, credit load, grade-point average, and SAT scores, one can use either a distance measure or a measure of association. For nominal variables such as gender, marital status, and race, one must use a similarity measure based on association coefficients ("matching-type" measures).

Distance Measures

The most popular distance measure used in cluster analysis, Euclidean distance, is the square root of the sum of the squared differences between corresponding measures. For example, if institutions are being classified according to enrollment, percent of undergraduate students, and graduation rate, then the distance measure is the square root of the difference between enrollments squared plus the difference in percent of undergraduates squared plus the difference between graduation rates squared.

Euclidean distance is one case of a more general formula for distances between points defined by multiple continuous variables, called the Minkowski metric and shown in equation (5.4).

$$d_{ij} = \left\{ \sum_{k=1}^{p} \left| X_{ik} - X_{jk} \right|^{r} \right\} \qquad [5.4]$$

For Euclidean distance, r =2. If r = 1, the distance measure is referred to as a city-block metric.

Another commonly used distance measure, Mahalanobis D^2, takes into account correlations among the predictors, as shown in equation (5.5), where $(X_i - X_j)$ is the vector of differences between corresponding scores and S is the variance-covariance matrix among measures.

$$D^2 = \left(\mathbf{X}_i - \mathbf{X}_j\right)' \mathbf{S}^{-1} \left(\mathbf{X}_i - \mathbf{X}_j\right) \qquad [5.5]$$

Because cluster analysis is based entirely on distances among objects according to the composite of criterion variables, one must be careful about the implications of using standardized versus unstandardized measures in computing these distances. That is, cluster analysis solutions will differ substantially when using unstandardized measures as compared to when using standardized versions of the same measures.

Matching-Type Measures (Association Coefficients)

The distance measures described to this point can only be used for continuous measures. In order to use nominal criteria, such as sector (public, private-non-profit, or proprietary) when clustering institutions, or gender and race/ethnicity when clustering students, a "matching-type" similarity measure, based on an association index, must be used.

For example, to generate a matching-type similarity measure for ethnicity, one would transform the single variable with, perhaps, six values (White, African American, Asian American, Hispanic, Native American, Non-Resident Alien, and Other) into five or six binary variables: White (0,1), African American (0,1), etc. One can either use one less variable than the number of values of the original variable, with the all zero's value representing one of the six values, or as many binary variables as valid values, saving the all zero's value for when race/ethnicity is missing.

To extend the example to multiple nominal criteria, consider comparing two people across a set of traits (e.g., gender, race/ethnicity, class level, major), where the numbers in each cell of

Person B	Person A	
	Has Trait	Does not have trait
Has Trait	A	b
Does not have trait	C	d

the following table represents the frequency of matches and mismatches on those traits.

For our example, assuming no missing values, there would be one variable representing gender, five variables for our race/ethnicity measure, and perhaps six variables to represent clusters of majors. For students who are of the same gender (female), race (Hispanic), and major cluster (social science) a=3, b=0, c=0, and d=10. It is important to note that this last component, d=10, reflects that two students who are identical on such criteria have more cases where they are common in terms of not meeting the criteria (they are both not male, not White, not African American, not Business majors, not Education majors, etc.) than where they meet the same criteria.

The numbers in this matching association example can be put together in various ways to derive a similarity coefficient. The "default" association would be calculated as (a+d)/(a+b+c+d). However, we may exclude the d value completely, and only consider when the pair have a trait in common relative to cases where one has a trait that the other doesn't (i.e., a/(a+b+c)). Alternatively, one could give double weight to the common traits relative to the rest (e.g., 2a/(2a+b+c+d)). As a heuristic algorithmic method, cluster analysis allows the user to choose among such measures according to which seem most logical for a given research question and data set.

Correlation Coefficients

Various forms of correlation coefficients, such as the Pearson product-moment correlation, or the Spearman rank-order coefficient, can also be used as a basis for determining similarity or difference among objects that are being considered for clustering. Traditionally, one uses correlation coefficients to examine the association between two measures across subjects (e.g., SAT scores and first-year GPA). In this case, however, the coefficient is calculated across the measures (criteria) and between two subjects.

Generating a Proximity Matrix

After selecting the variables and distance/similarity measure, the next step is to produce a matrix, variously called the distance or proximity matrix, which contains the composite distance/similarity measure for each and every pair-wise combination of objects. The proximity matrix contains a row and column for each object (person or thing) and the cells represent the distance or similarity measure between each pair. Thus the distance matrix is symmetric and has a diagonal of zeroes for distance matrices, or ones for similarity measures.

For one common institutional research clustering application—selecting peer institutions—this is the final step in the analysis. The row or column containing the target institution can be isolated and examined to see which other institutions are "closest" or "most similar." An example of this technique is provided in the following section.

Choosing a Clustering Technique

If finding the "nearest neighbors" for a single institution is not the final goal, one would then proceed to the final stage of a cluster analysis, wherein an algorithm is chosen for determining cluster membership. There are two general classes of algorithms—hierarchical and partitioning—and several methods within each class.

Hierarchical Algorithms

Hierarchical algorithms operate in one of two directions. Agglomerative methods start with each object in its own cluster and then put closest items and clusters together into larger groupings until some termination criteria is reached. Divisive algorithms start with all points together in a single cluster and then partition the objects into smaller groups by splitting off items and small clusters.

Agglomerative methods differ according to the criteria by which distance between clusters is determined. That is, once there is more than one object in a cluster, one must decide how to determine the distance between that cluster and another object (unclustered point) or cluster. The Single linkage (or Nearest Neighbor) technique calculates distance according to the closest points in two clusters or the distance between an unclustered object and the closest point in the cluster. Complete linkage (Furthest Neighbor) calculates distance according to the largest distance among clusters (points farthest away from each other). Average linkage, as its name implies, bases distance on the average of all distances between points when evaluating point-cluster or cluster-cluster distances. Ward's Error Sum of Squares is another popular method for calculating distances within a hierarchical agglomerative cluster analysis. It uses an objective function that minimizes the sum of squared deviations between the points and the cluster mean. This method provides some basis (i.e., the objective function) for judging when to stop forming clusters.

Divisive hierarchical clustering methods, which start with one group of the whole and then partition objects into smaller clusters, require the researcher to choose a technique for determining which points or clusters to detach from the larger group. The Splinter-Average Distance technique first removes the object that has the greatest average distance from all other objects. Each object is tested to see if it is closer on average to the original group or the splinter group. Once all objects are assigned to one of the two groups, the process is repeated on each group.

Decision tree clustering methods are also considered to be within the hierarchical divisive camp. However, decision trees depart from the methods so far described in several fundamental ways and will be considered separately.

Partitioning Algorithms

Partitioning algorithms start with an a priori determination of the number of clusters. Criteria are then selected for optimizing distances among clusters.

Although partitioning algorithms require prior statement of the number of final clusters, some methods allow cluster numbers to vary during the course of analysis. The methods of partitioning generally differ with regard to how the clusters are initially determined, how objects are assigned to clusters, and how objects are reallocated to clusters.

K-Means clustering is the most popular partitioning algorithm. The researcher specifies the number of clusters (K) and has the option of specifying the initial location of each cluster center (according to a composite of the criterion variables). If the researcher does not specify the initial cluster centers, the algorithm chooses locations that are distributed evenly throughout the data space. The algorithm then adjusts the cluster locations as to minimize the Euclidean distance between the objects and cluster means.

Trace-based methods are a class of partitioning methods that either maximize between-group dispersion or minimize within-group dispersion using various combinations of the trace and determinants of the matrices representing each of these components.

Cluster Analysis Examples

As mentioned above, one of the most common institutional research applications of cluster analytic techniques—identifying peer institutions—typically focuses on the row or column in a proximity matrix pertaining to the target institution. As such, the analysis never actually employs a clustering algorithm. For our current example we will first run a set of institutions through a clustering algorithm and then examine the use of the proximity matrix to find a target institution's "nearest neighbors," comparing the results of the two processes.

The data for this example come from the IPEDS data sets collected annually by the National Center for Education Statistics. Table 13 lists the data elements employed in this analysis. This analysis will only consider the 144 Doctoral Extensive universities (according to the 2000 Carnegie Classification system) that have complete data on all these variables.

For our initial analysis, we will employ a hierarchical, agglomerative algorithm on the standardized (Z-score) values of the criteria, using a Euclidean distance measure and the between-groups (average-linkage) method for determining distance between points and cluster. We will also request solutions that produce five through eight clusters.

The output of a cluster analysis procedure tends to be very voluminous and not very informative. The Agglomeration Schedule, for example, lists the cluster numbers being combined at each stage, although it is unknown which points are in which clusters. The Proximity Matrix can also be requested, but the printed form is generally burdensome compared to the saved data set form described below. A Dendrogram or Icicle Plot can be requested, showing which points are being put together or separated at each stage. With more than twenty or thirty cases, the output is excessive. The key result for the

Table 13
Variables from NCES IPEDS Data Sets Used for Peer Institution Identification and Clustering Example

Variable	Label
totenr	Enrollment - Total Headcount
ftpct	Enrollment - Percent Full-Time
wompct	Enrollment - Percent Women
ugpct	Enrollment - Percent Undergraduate
minpct	Enrollment - Percent Minority
pbus	Degrees - Percent Business, Management, Marketing and Related Support Services
peduc	Degrees - Percent Education
pengin	Degrees - Percent Engineering
phealth	Degrees - Percent Health Professions and Related Clinical Sciences
resexp	Expenditures - Research
totexp	Expenditures - Total Operating
pontrack	Faculty - Percent Tenured or On-Track
ptenure	Faculty - Percent of On-Track who are Tenured
satact	Average SAT or ACT Equivalent
price	Total Price (In-State, On-Campus)

cluster analysis is the assigned cluster membership, which can be saved onto the data set for each requested solution.

Another important result for our efforts to identify "Nearest Neighbors" is the proximity matrix. Unfortunately, when using the graphic user interface dialogues in SPSS, the proximity matrix is created, used, and deleted, "behind the scenes." The only way to capture the proximity matrix is through the use of syntax. Specifically, the following syntax produces the proximity matrix for our example:

```
Proximities  totenr ftpct wompct ugpct minpct pbus
peduc pengin phealth resexp totexp pontrack ptenure
satact price
    /MATRIX OUT ('C:\temp\prox.sav')
    /VIEW= CASE
    /MEASURE= SEUCLID
    /ID= INSTNM
    /STANDARDIZE= VARIABLE Z
    /PRINT NONE.
```

The proximity matrix created by this syntax would be saved into an SPSS dataset as "c:\temp\prox.sav," with a row and a column for each of the 144 institutions in this analysis.

Table 14 shows the number of institutions placed into each cluster in the five, six, seven and eight cluster solutions produced by the hierarchical agglomerative analysis. The results show that the vast majority of institutions were placed in the first cluster, with one or two institutions split off to form most of the remaining clusters, until the eight cluster solution, when a group of thirty forms one of the other clusters.

Table 14
Cluster Frequencies for the 5 through 8 Cluster Solutions

Cluster	5 Cluster Solution	6 Cluster Solution	7 Cluster Solution	8 Cluster Solution
1	138	137	137	107
2	1	1	1	1
3	2	2	1	30
4	1	1	1	1
5	2	2	1	1
6		6	7	8
7			1	2
8				1

As an alternative, the "Two-Step" cluster method within SPSS can be used to first identify the optimal number of clusters before performing the analysis. In this case, the only solution provided was a one-cluster solution (perhaps suggesting some validity to the Carnegie Classification system). However, using the K-Means partitioning method forces the cluster membership to be more uniform. Table 15 shows the number of institutions in each cluster for the five through eight cluster solutions using the K-Means method. Although more evenly distributed, the cluster sizes still vary greatly. More importantly, if a target institution is located in one of the small clusters, the solution would not be suitable for peer institution determination.

Table 15
Cluster Frequencies for the 5 through 8 Cluster Solutions

Cluster	5 Cluster Solution	6 Cluster Solution	7 Cluster Solution	8 Cluster Solution
1	25	26	7	7
2	75	78	54	54
3	12	3	30	3
4	3	20	3	2
5	29	7	3	40
6		10	40	30
7			7	7
8				1

As mentioned above, the proximity matrix is usually far more useful for determining the "nearest neighbor" institutions to a target institution. For demonstration purposes, we will choose the University of Southern California, which falls within the thirty institution cluster three of the hierarchical eight cluster solution and also within the thirty institution cluster six of the K-Means eight-cluster solution. Table 16 shows the twenty-nine nearest neighbors to

Table 16
Nearest Neighbors Compared to Hierarchical and K-Means Solutions

Institutions	Standard Euclidean Distance	Rank of Distance	Hierarchical 8 Cluster	K-Means 8 Cluster
UNIVERSITY OF SOUTHERN CALIFORNIA (Target)	0	0		
NORTHWESTERN UNIVERSITY	6.0301	1	X	
BOSTON UNIVERSITY	7.7255	2	X	X
WASHINGTON UNIVERSITY IN ST LOUIS	8.3194	3	X	X
GEORGETOWN UNIVERSITY	9.6644	4	X	
EMORY UNIVERSITY	10.3252	5	X	
UNIVERSITY OF MIAMI	11.3731	6		X
TULANE UNIVERSITY OF LOUISIANA	11.6378	7	X	
GEORGE WASHINGTON UNIVERSITY	11.7293	8	X	
UNIVERSITY OF PITTSBURGH-MAIN CAMPUS	11.7771	9		X
NEW YORK UNIVERSITY	11.8242	10	X	
VANDERBILT UNIVERSITY	13.0488	11	X	
UNIVERSITY OF MARYLAND-COLLEGE PARK	13.5931	12		X
UNIVERSITY OF ILLINOIS AT CHICAGO	14.5512	13		X
UNIVERSITY OF PENNSYLVANIA	14.6068	14	X	
UNIVERSITY OF VIRGINIA-MAIN CAMPUS	14.8273	15		X
DUKE UNIVERSITY	15.1560	16	X	
CASE WESTERN RESERVE UNIVERSITY	15.9863	17	X	
NORTHEASTERN UNIVERSITY	16.1986	18	X	
UNIVERSITY OF IOWA	16.3627	19		X
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN	16.4958	20		X
SUNY AT STONY BROOK	16.9441	21		X
UNIVERSITY OF WASHINGTON-SEATTLE CAMPUS	16.9477	22		
UNIVERSITY OF CINCINNATI-MAIN CAMPUS	16.9592	23		
MARQUETTE UNIVERSITY	17.0076	24	X	
SUNY AT BUFFALO	17.2150	25		
BOSTON COLLEGE	17.3531	26		
UNIVERSITY OF CHICAGO	17.4265	27	X	X
UNIVERSITY OF FLORIDA	18.3060	28		X

Table 16 (continued)
Nearest Neighbors Compared to Hierarchical and K-Means Solutions

Institutions	Standard Euclidean Distance	Rank of Distance	Hierarchical 8 Cluster	K-Means 8 Cluster
UNIVERSITY OF WISCONSIN-MADISON	18.9855	29		
UNIVERSITY OF KENTUCKY	19.7626	30		X
TEMPLE UNIVERSITY	19.9824	31		X
UNIVERSITY OF MISSOURI-COLUMBIA	20.0212	32		X
CARNEGIE MELLON UNIVERSITY	20.4008	34	X	
TUFTS UNIVERSITY	20.5449	35	X	
UNIVERSITY OF CALIFORNIA-BERKELEY	20.6761	37		X
CORNELL UNIVERSITY-ENDOWED COLLEGES	21.3958	40	X	X
UNIVERSITY OF ARIZONA	22.3728	42		X
THE UNIVERSITY OF TEXAS AT AUSTIN	22.5967	44		X
UNIVERSITY OF ROCHESTER	22.7624	46	X	X
UNIVERSITY OF UTAH	22.8306	47		X
MICHIGAN STATE UNIVERSITY	23.5501	53		X
UNIVERSITY OF CALIFORNIA-IRVINE	23.7519	54		X
TEXAS A & M UNIVERSITY	24.1569	57		X
PURDUE UNIVERSITY-MAIN CAMPUS	24.2557	59		X
HARVARD UNIVERSITY	24.5868	62	X	
UNIVERSITY OF NOTRE DAME	24.8950	66	X	
UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL	25.9354	69		X
SAINT LOUIS UNIVERSITY-MAIN CAMPUS	26.7437	75	X	
COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK	26.7461	76	X	
RICE UNIVERSITY	28.4091	81	X	
LEHIGH UNIVERSITY	29.3976	87	X	
PRINCETON UNIVERSITY	32.3095	99	X	
YALE UNIVERSITY	32.4000	100	X	X
UNIVERSITY OF NEW MEXICO-MAIN CAMPUS	32.7737	103		X
UNIVERSITY OF ALABAMA AT BIRMINGHAM	40.9372	125		X
GEORGIA INSTITUTE OF TECHNOLOGY-MAIN CAMPUS	41.4375	127	X	
YESHIVA UNIVERSITY	48.4981	136	X	

the University of Southern California according to the proximity matrix, followed by other institutions identified through either the hierarchical or partitioning derived clusters.

The variation between cluster solutions is quite evident in Table 16. Fifteen of the twenty-nine institutions (aside from USC) that occupy the cluster three of the eight-cluster hierarchical solution and twelve of the twenty-nine from cluster six of the eight-cluster K-Means solution are among the twenty-nine nearest neighbors derived from the proximity matrix. However, only three of these institutions are in common to all three solutions (and only four institutions are in common between the two cluster groups). Institutions that are in the cluster solutions but not among the nearest neighbors are shown below the twenty-ninth institution in the proximity matrix (i.e., below the University of Wisconsin, Madison). Many of these institutions are relatively close in distance to the target, including six of the eight next nearest neighbors. However, ten of the institutions appearing in one of the two cluster solution groups are in the bottom half (i.e., below rank seventy-two) in distance from the target institution.

The results of this cluster analysis highlight the instability of cluster analytic methods as a "pattern recognition" algorithm. For this reason, most peer institution techniques have relied on the proximity matrix for identifying nearest neighbors. It is important to remember, however, that the proximity matrix will produce markedly different neighbors depending on what measures are employed, and whether the analysis is based on standardized or unstandardized measures.

Cluster Analysis References

Institutional Research Applications

Cowles, D. & Franzak, F. (1991). Divide and conquer: Applying the marketing concept of 'segmentation' to the placement function. *Journal of Career Planning and Employment, 51*, 59-63.

Goldgehn, L. A. (1989). Admissions standards and the use of key marketing techniques by United States colleges and universities. *College and University, 65*, 44-55.

Muffo, J. A. (1987). Market segmentation in higher education: A case study. *Journal of Student Financial Aid, 17*, 31-40.

Rickman, C. A. & Green, G. (1993). Market segmentation differences using factors of college selection. *College and University, 68*, 32-37.

Wakstein, J. (1987). Identifying market segments. In R. S. Lay and J. J. Endo (eds.), Designing and using market research, *New Directions for Institutional Research*, No. 54. San Francisco: Jossey-Bass.

General References

Aldenderfer. M. S. & Blashfield, R. K. (1984). Cluster analysis. *Quantitative Applications in the Social Sciences, No. 44.* Beverly Hills, CA: Sage.

Blashfield, R. K. and Aldenderfer, M. S. (1978), The literature on cluster analysis, *Multivariate Behavioral Research, 13*, 271 -295.

Dillon, W. R. & Goldstein, M. (1984). *Multivariate analysis: Methods and applications.* New York: Wiley.

Klastorin, T .D. (1983), Assessing cluster analysis results, *Journal of Marketing Research, 20*, 92 -98.

Marriott, F. H. C. (1971), Practical problems in a method of cluster analysis, *Biometrics, 27*, 501 -514.

Mezzich, J. E and Solomon, H. (1980), *Taxonomy and behavioral science*, New York: Academic Press.

Milligan, G. W. and Cooper, M. C. (1985), An examination of procedures for determining the number of clusters in a data set, *Psychometrika, 50*, 159 -179.

Sokal, R. & Sneath, P. (1963). *Principles of numerical taxonomy.* San Francisco: W. H. Freeman.

Decision Trees

Decision trees are a form of hierarchical, divisive cluster analysis because the analysis begins with all objects belonging to a single cluster with subgroups subsequently split off. However, decision trees differ from the clustering methods described thus far in two fundamental ways:

1. A single criterion (outcome) measure is used to maximize differences among groups (e.g., students could be clustered to maximize differences between those who are retained and those who are not retained; institutions could be differentiated based on maximizing differences in research expenditures).
2. Clusters are partitioned using one variable at a time, with subsequent sub-clusters determined separately for each "branch" of the tree using any of the remaining variables.

Because of the use of a criterion outcome, decision trees are somewhat of a 'hybrid' (in function) between clustering and discriminant analysis. That is, the criterion outcome variable does not define the groups, as in discriminant analysis, but the groups are defined to maximize differences according to the criterion.

For example, consider student retention. A discriminant analysis would generate a linear combination of the classification variables that best distinguishes returning from non-returning student groups. The decision tree would find the best interaction of values among classification variables that produces groups with the greatest differences in retention rates. The resulting groups would be defined in terms of the classification variables (e.g., white, females, over 25 years old, for which the group retention rate is xx%) and not the criterion variables (e.g., returning students, who tend to have higher

proportions of white females over twenty-five years of age compared to the non-returning group).

Decision trees are particularly useful for identifying which variables distinguish best among groups and for formulating group membership prediction rules. These membership rules are then helpful for predicting future observations. Moreover, the illustration of these differences through the graphical representation employed by decision trees is very useful for disseminating the results in a relatively non-technical format.

A range of decision tree techniques is available. The SPSS® AnswerTree™ software includes three decision tree algorithms that all use a "brute force" method to examine how well each and every possible classification variable partitions objects based on the criterion variable. The three algorithms—Chi-Square Aided Interaction Detection (CHAID), Clustering and Regression Trees (C&RT), and Quick, Unbiased, Efficient, Statistical Tree (QUEST)—perform several common functions:

- Merge categories of the predictor variables so that non-significantly different values are pooled together
- Split the variables at points that maximize differences
- Stop branching when further splits do not contribute significantly
- Prune branches from an existing tree
- Validation and error estimation

The first of these features can be useful in and of itself as a pre-cursor to other analyses. The merging of categories of predictors, sometimes called discretization, helps determine appropriate cutoff points for predicting outcomes. For example, many college admissions offices employ an admissions index made up of a combination of entry characteristics, such as entrance exam scores (SATs or ACTs), high school grades, co-curricular activities, and so on. The discretization function of a decision tree analysis determines the cutoff points that maximize differences on an outcome measure, such as first-year college grades or retention to the second year. This can be a very useful method for determining cutoff points for admittance, or for validating existing cutoffs.

The three decision tree algorithms included in AnswerTree™ also have some fundamental differences. The C&RT and QUEST algorithms are relatively similar and employ only binary splits. That is, at each point in the tree, the group can only be split in two. The CHAID algorithm allows for multi-group splits and so provides greater flexibility for typical institutional research applications. The C&RT and QUEST algorithms offer several advantages for financial applications and so are commonly used for loan and insurance eligibility.

Further details on the features and suitability of these techniques are available in the references provided below. However, the best way to illustrate

the uses of a decision tree analysis is by example. In the following section, we examine a CHAID decision tree that segments the Doctoral Extensive institutions to maximize differences in average research expenditures.

CHAID Decision Tree Example

Our example decision tree analysis employs the same institutional data set used for the cluster analysis with the addition of two nominal variables: institutional control (public or private) and whether the institution offers a medical degree (yes or no). These two variables were chosen to demonstrate how CHAID can employ both nominal and interval/ratio measures. Moreover, an institution's research expenditures are impacted greatly by the presence of a medical school.

When reading the data set into the Answer Tree software, the user must specify a "target variable." This is the outcome criterion that will be used to maximize differences on all other (predictor) variables for group identification. In our example, annual research expenditures is chosen as the target variable. When the data set is brought into the interactive CHAID procedure, the software initially presents a "root node" which shows summary statistics for the target variable. For this example, the root node shows that the average research expenditures for the 144 institutions are just shy of $165 million. The root node also notes that the standard deviation is just shy of $158 million.

The user can proceed with building the CHAID decision tree in one of two general ways: automatically, letting the software choose the best predictors at each level; or manually, letting the user choose which variable to enter at each level and sub-level of the tree. Regardless of which method is chosen, the user can subsequently "prune" and re-grow branches of the tree based on automatic or manual selection.

Figure 1 shows the default decision tree produced for this analysis. It includes only one branching variable—total operating expenditures—which parsed the institutions into two groups. Specifically, the cutoff on total operating expenditures of just over $707 million split the institutions into one group of eighty-six institutions that averaged just under $73 million in research expenditures and one group of fifty-eight institutions that averaged just over $300 million.

The use of total operating expenditures to distinguish among institutions according to research expenditures is not very revealing, because research expenditures is a portion of total expenditures. Therefore, the user might wish to see what other variables account for significant differences between institutions on this criterion. Figure 2 shows a portion of the predictor selection table, revealing that only three other variables will generate statistically significant group differences in research expenditures given the current "rules:" whether the institution has a medical school, the percent of women enrolled, and the percent of underserved minority populations enrolled.

Figure 1
Default CHAID Decision Tree Distinguishing Doctoral Extensive Universities on the Research Expenditures Criterion

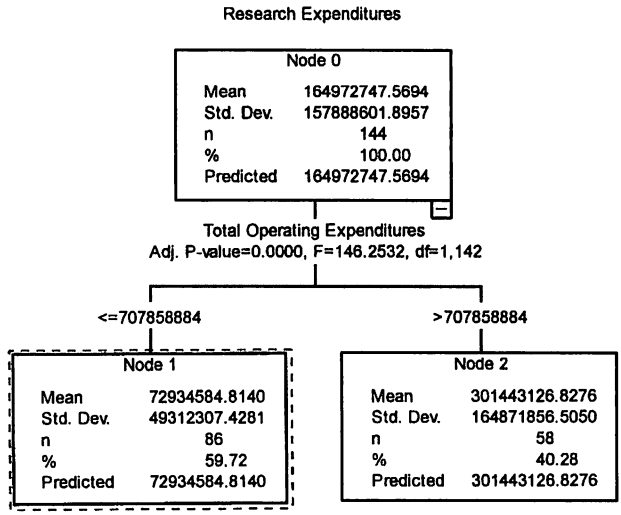


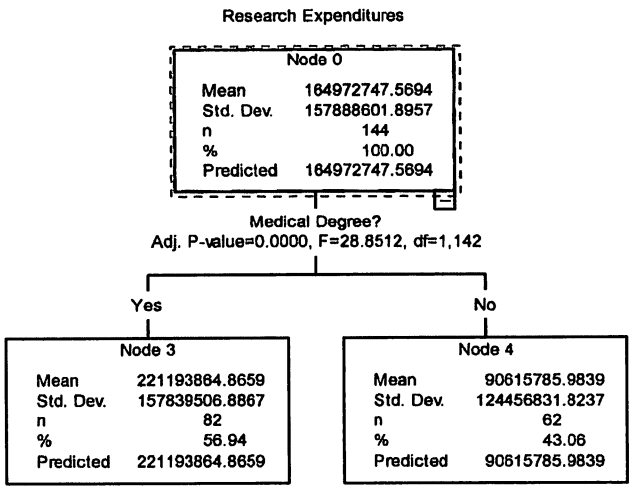
Figure 2
Predictor Selection Table for CHAID Analysis of Institutional Research Expenditures

Predictor	Nodes	Split Type	F	D.F.	Adj.Prob.
Total Operating Exp...	2	Default	146.2532	1, 142	0.000000000
Medical Degree?	2	Default	28.8512	1, 142	0.000000312
Enr Pct Women	2	Default	24.3031	1, 142	0.000020385
Enr Pct Underserve...	2	Default	1.2051	1, 142	1.000000000
Institution Name	*	Arbitrary	*	*	*
Total Enrollment	*	Arbitrary	*	*	*
Enr Pct Full-Time	*	Arbitrary	*	*	*
Enr Pct Undergradu...	*	Arbitrary	*	*	*
Degrees Pct Busine...	*	Arbitrary	*	*	*
Degrees Pct Educa...	*	Arbitrary	*	*	*

Choosing the medical school field as the predictor results in the Tree shown in Figure 3. This tree shows that the eighty-two universities that offer a medical degree had average research expenditures of just over $221 million compared to just over $90 million for the sixty-two universities that offer no medical degree.

With this tree, as well as with the prior tree showing total operating expenditures as the strongest predictor of research expenditures, no further branching will occur under the default stopping rules. However, the user can alter these rules, which include statistical significance criteria (i.e., p-level),

Figure 3
CHAID Decision Tree Using Medical Degree Status to Distinguish Doctoral Extensive Universities on the Research Expenditures Criterion

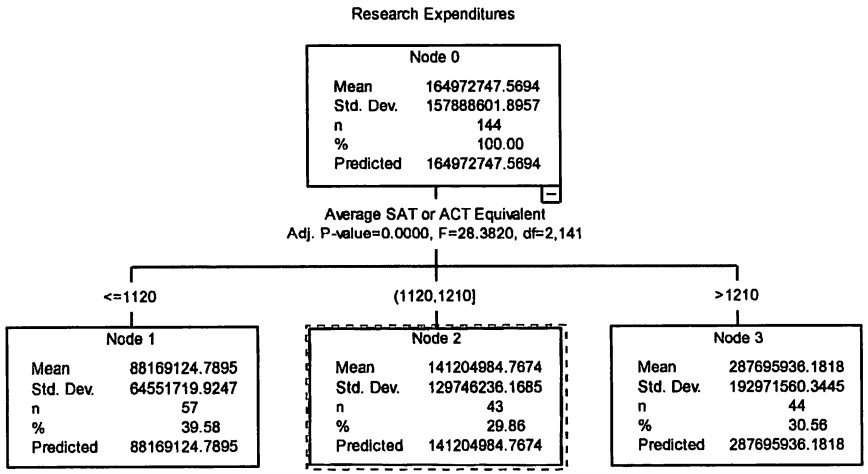


a default depth of three levels, and a minimum of fifty observations per group. In this case, for example, if the minimum group size is dropped to twenty, the Entering Student SAT/ACT variable becomes the second strongest predictor (after total operating expenditures) resulting in the tree shown in Figure 4.

The model shown in Figure 4 distinguishes between three groups of universities. The group of fifty-seven institutions with the lowest average entry scores (SAT or ACT equivalent less than 1120) averages just over $88 million in research expenditures. The middle group of forty-three institutions with average entry scores between 1120 and 1210 averages just over $140 million in research expenditures. Finally, the forty-four institutions for which the average entry score of students is above 1210 average about $287 million in research expenditures.

So far we have considered only single node trees. Figure 6 shows an extension of the Figure 5 model. Specifically, the "low average entry score" institutions have been further divided into the twenty-nine among them that have medical schools—with average research expenditures of around $127 million and the twenty-eight schools among them that do not have medical schools, which average research expenditures approximately $48 million. The "middle entry score" group has been divided into three subgroups based on a different variable: total enrollment. That is, the forty-three original institutions were divided into eleven relatively small institutions (enrollments less than 15,093) that average $37 million in research expenditures, fifteen mid-sized institutions (total enrollment between 15,093 and 26,552) that

Figure 4
CHAID Decision Tree Using Entering Student Average SAT/ ACT Score to Distinguish Doctoral Extensive Universities on the Research Expenditures Criterion



average $89 million in research expenditures, and the seventeen relatively large institutions (total enrollment greater than 26,552) that average $255 million in research expenditures. The third node of the original model was not divided any further.

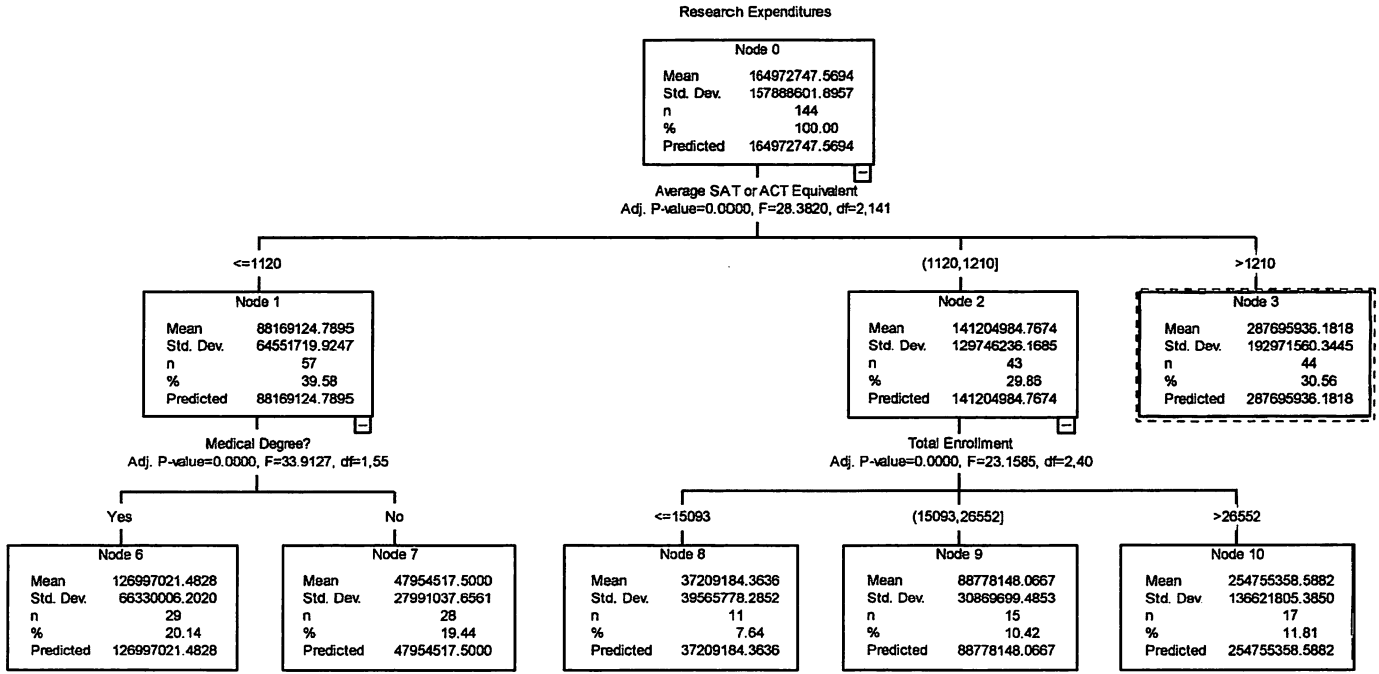
This last model shows some unique and powerful features of decision trees in general and CHAID in particular. This technique lets the user explore different combinations of variables to predict differences in the outcome criterion. Moreover, the user can choose different "interaction" variables within levels of the tree, as we saw by using the medical school distinction for one node at the second level, and using total enrollment for a different node. Decision trees can be a very powerful tool for exploratory analysis of group differences. However, because they are so data driven (i.e., potentially unstable for different samples), the user should be careful to test models across multiple samples. In addition, exploratory findings should be subject to confirmatory statistical analyses to determine if they are reliable.

Decision Tree References

Institutional Research Applications

Lay, R. S. & Maguire, J. J. (1983). Computer aided segmentation analysis: New software for college admissions marketing. *Journal of College Admissions, 101*, 32-36.

Figure 5
Further Predictors for CHAID Decision Tree Model



Thomas, E. H. & Galambos, N. (2004). What satisfies students? Mining student-opinion data with regression and decision tree analysis. *Research in Higher Education, 45*, 251-269.

General References

Fleischer, R., (1999). Decision trees: old and new results. *Information and Computation, 152*, 44.

Kamber M, Winstone L, Gong W, Cheng S, Han J (1997). Generalisation and decision tree induction: Efficient classification in data mining. In *Proceedings of 1997 International Workshop on Research Issues on Data Engineering (RIDE'97)* Birmingham, England, 11-120.

Liu B, Xia Y, & Yu P (2000). Clustering through decision tree construction. In *Proceedings of the ACM International Conference on Information and Knowledge Management,* Washington, DC.

Skinner, D. C. (1999). *Introduction to decision analysis (2nd edition).* Sugar Land, TX: Probabilistic Publishing.

Sonquist, J. A. & Morgan, J. N. (1964). The detection of interaction effects. *Monograph No. 35*, Institute for Social Research, University of Michigan.

Choosing between Cluster Analysis and Decision Trees

Borden, V. M. H. (1995). Segmenting student markets with a student satifaction and priorities survey. *Research in Higher Education, 36*, 73-88.

Summary and Conclusions

This chapter presented two techniques for examining differences among pre-existing groups (discriminant analysis and logistic regression), and one general class of analyses (cluster analysis) along with several specific manifestations (hierarchical and partitioning techniques; use of the proximity matrix to identify nearest neighbors; and CHAID decision trees for mining complex interactions among predictors of group outcomes) for identifying groups that do not exist, a priori. Despite its length, this chapter has only scratched the surface on each of these topics. References are provided at the end of each section to guide the reader who seeks more in-depth information about each technique and its application to some typical institutional research problems. After gaining sufficient general understanding, the author suggests that the reader consider developing more in-depth expertise and experience in three areas.

First, logistic regression has become the coin of the realm for examining differences in existing groups. With the extension to multi-group (polychotomous) outcomes available through multinomial logistic regression and to rank-order outcomes, through ordinal logistic regression, this family of related techniques provides sufficient scope to cover virtually any analysis of pre-existing groups that an institutional researcher would likely encounter.

The second recommended focus is on the selection of similarity and distance measures and the use of the proximity matrix to identify nearest neighbors. The example in this chapter employed only interval/ratio measures to develop the proximity matrix for peer institution selection. However, one might also want to include a set of categorical variables, such as control, geographic location, or the presence or absence of particular programs or facilities. Matching-type measures, for example, enable one to combine measures on varying scales (nominal, ordinal, interval, and ratio) into a single similarity measure.

Finally, decision trees, and CHAID in particular, are very useful techniques for exploring relationships among objects, especially when one has a meaningful criterion measure. In our example, we explored group differences among research universities with research expenditures as the criterion. CHAID analyses are also useful in examining the efficacy of admissions criteria in relationship to a success-related outcome, such as freshman year grades or retention to the second year. However, as an exploratory technique, it is important to conduct more confirmatory statistical analyses on the findings to establish their reliability at a more general level.

The methods considered in this chapter provide institutional researchers with an array of useful tools for many common IR applications. In closing, it is important to underscore the need for well-grounded conceptualization and theory to render these powerful tools and techniques useful. Like any power tool, they can produce wonderful results if the user has a well-conceived design and is skilled in the use of all available features. Although safety glasses are not required, the repercussions for irresponsible use can be just as damaging.

Endnotes

[1] At this point we are considering only "dichotomous" group outcomes as handled by the t-statistic. We will touch briefly upon "polychotomous" (more than two group value) outcomes later.

[2] When using this technique, an adjustment must be made in the error terms used to test the coefficients in the second stage model, as proposed by Heckman (1979).

[3] The null population value for the odds ratio (i.e., the exponent of b) is one, since $e^0 = 1$.

Chapter 6
Applied Multivariate Statistics

Mary Ann Coughlin

In layperson's terms, multivariate statistics are any inferential statistical procedures that utilize one or more indicator or predictor variables to explain multiple outcome variables. Multivariate statistics allow us to analyze complex data sets. Further, these statistics provide for analysis where many independent variables are used and multiple dependent variables are present that correlate to each other to varying degrees. While whole volumes have been written on this topic alone, to date, little has been written on the direct application of these statistics within the field of institutional research. Within the context of this chapter, my hope is to introduce the area of multivariate statistics as applied in institutional research. The chapter is designed to provide the reader with underlying theories of multivariate statistics and direct applications of these statistical procedures to problems commonly faced in institutional research. The goal of this document is to present a summary of three multivariate techniques (path analysis, exploratory factor analysis, and confirmatory factor analysis) and to demonstrate each with examples and illustrations. Additionally, the chapter will end with a brief introduction to structural equation modeling. Given the breadth of the material covered, please consult the references provided for depth greater than appropriate from a document of this scope. The chapter will be organized into the following sections: path analysis, factor analysis, and structural equation modeling.

Path Analysis

As we discuss our first multivariate statistical procedure, we will review the theoretical and logical associations that exist between path analysis and techniques presented in prior chapters. Next, we will describe this technique in depth through the use of a case study application of path analysis in institutional research. We will end this section with a summary of the interpretation of our case study and a discussion of the application of this technique in institutional research.

Statistical and Theoretical Background

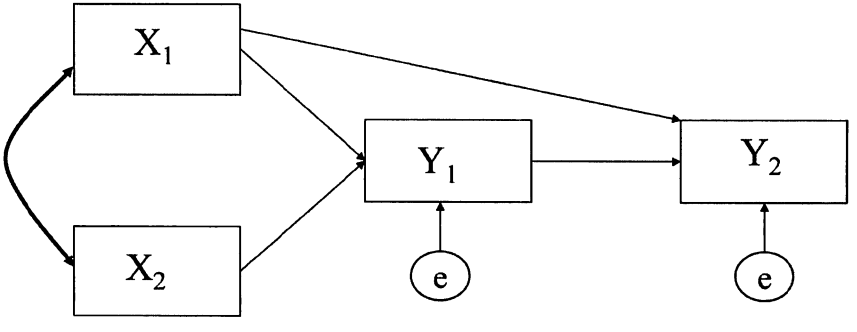
As was identified in prior chapters, the association between two variables is the statistical basis for many inferential statistical techniques. Correlation measures both the strength and the direction of the relationship between two variables. The relationship that is measured between the variables is the mutually shared relationship between the two variables, thus no independent or dependent variables are identified. As a result no causality can be implied from that shared relationship. Bivariate regression logically extends the concept

of correlation by using a linear predictive model to determine the degree to which a single predictor (i.e., independent) variable can be used to predict or explain a criterion (i.e., dependent) measure. Multiple linear regression can be described as a logical extension of bivariate regression where multiple predictor variables are used to create a more complete explanation of the one criterion measure. In almost all cases, multiple regression is considered to be advantageous as seldom can any outcome be adequately explained by a single predictor variable. Our first multivariate technique, path analysis can logically be viewed as an extension of multiple linear regression. In path analysis, multiple predictor variables are used in a linear model to predict or explain multiple criterion measures. A focus in path analysis is the predictive ordering of variables; thus, path analysis allows the researcher to test an integrated theory of influences among a set of variables. The theoretical similarities and differences between regression and path can be summarized by stating that: the model 'X predicts Y' is a regression model, whereas, 'X predicts or influences Y and Y predicts or influences Z' is a path analysis model.

Path analyses are commonly displayed in figures (path diagrams) that represent the relationships that are being tested in the model. Path diagrams are drawn so that the flow of influence of the variables in terms of causal ordering is from left to right. Figure 1 displays a theoretical path model that describes the relationship between two predictor variables and two criterion or outcome measures. As we move from regression analysis to path analysis and review this figure, we must clarify new terms. The path diagram displays the observed variables and the proposed relationships between the variables. Observed variables are those variables that are 'observed' or measured by the researcher. These variables are displayed as squares on the path diagram and arrows display the relationships between the variables.

In path analysis, all variables are observed variables and we label variables as being exogenous or endogenous. Endogenous variables are those

Figure 1
Theoretical Path Model



variables that the model attempts to predict or explain. Arrows point toward endogenous variables. In our theoretical model displayed in Figure 1, Y_1 and Y_2 are endogenous variables, because they are being explained within the model. Thus, endogenous variables are similar to dependent or outcome measures. Exogenous variables are those variables that the model makes no attempt to predict or explain. Variables that do not have arrows directed toward them are exogenous. In Figure 1, X_1 and X_2 are exogenous; no explanation exists for their variance within the model and no arrows are pointing to these variables. The exogenous variables are by definition similar to predictor variables. So while endogenous variables are similar to dependent variables and exogenous variables are similar to independent variables, the change in terminology must be made in path analysis because variables may be used as both independent and dependent variables within a particular path analysis. In Figure 1, for example, Y_1 is an endogenous variable that is predicted from X_1 and X_2, but Y_1 is also used as an independent variable to predict our second endogenous measure Y_2. In path analysis, we use the terms exogenous and endogenous as opposed to independent or dependent variables.

Also in path analysis, we identify the influences of variables within the model as either direct or indirect effects. Direct effects are those parameters or coefficients that estimate the 'direct' influence one variable has on another. The lines on the path diagram indicate the direct effects or casual relationships that are identified in the model. In Figure 1, X_1 and X_2 have a direct effect on Y_1; in addition, X_1 and Y_1 have a direct effect on Y_2. Indirect effects are not indicated on the path diagram. In our model, X_1 and X_2 have an indirect effect on Y_2 through Y_1.

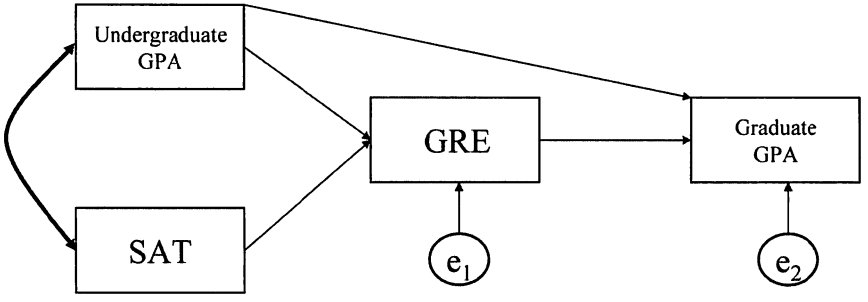
Relationships between exogenous variables are commonly displayed on the path diagram as curved double-headed arrows. As is the case in any prediction model, errors in prediction exist. The errors in prediction are normally indicated on the path diagram as circles that contain an "e," as they are estimates of the errors in prediction and not observed variables. Sometimes because of printing constraints or to keep a path diagram from appearing as cluttered, the double-headed arrows or the arrows and coefficients for the residuals may be omitted from the diagram. In these cases, the terms are assumed to be present in the model, although not displayed. Now let us explore path analysis through a case study application in institutional research.

Case Study: An Examination of Performance in Graduate School

This case study will explore path analysis by examining the ability of student characteristics to predict performance in graduate school. The dean of the graduate program at a small public university is interested in determining the extent to which student characteristics can be used to predict performance in graduate school. The dean is concerned that students are not being properly identified for selection into her graduate program. The dean is concerned that

too much emphasis is being placed on standardized tests and not enough emphasis is being placed on prior academic achievement. After identifying a model that will explore the ability of SAT and undergraduate Grade Point Average (GPA) to predict Graduate Record Examination (GRE) scores and the ability of these variables to predict graduate school GPA, path analysis was identified as the appropriate statistic to analyze the data. Path analysis is the appropriate statistical analysis for this research question, because of the presence of multiple dependent variables and the ordering of variables across time. SAT and undergraduate GPA will be used as the predictor variables to explain the variance of GRE scores. Additionally, in the second stage of this model, the researcher will test the ability of these variables to predict graduate school GPA. Figure 2 graphically displays this model. In this model, we have two endogenous variables, GRE and graduate school grade point average, and two exogenous variables, SAT and undergraduate grade point average.

Figure 2
Case Study: Path Diagram



Before exploring our case study further, we should first review the basic statistical assumptions of path analysis and ensure that our data meet the requirements of the statistical procedure. As discussed earlier path analysis is based on the techniques of multiple regression; as a result the assumptions of multiple linear regression described in chapter three apply for path analysis as well. The statistical assumptions regarding measurement error and the multicollinearity of predictor variables again warrant the attention of the researcher. Modest levels of multicollinearity are, however, tolerated within path analysis. While, model specification errors are problematic in regression analysis, they are particularly worrisome in path analysis because of the possible additive and interactive influences of these errors in the prediction of multiple endogenous variables. Specification errors refer to those errors made in correctly identifying the regression or path model. The researcher could either incorrectly include a variable in the model that does not belong, or may omit a variable from the model that does belong. The latter specification error is more problematic than the former, but as was identified earlier in the

regression chapter of all of the problems facing analysts when using predictive techniques, this error is perhaps the most difficult to overcome. Virtually all models are subject to criticism for omitting relevant factors. Finally, a question arises about how large a sample size is sufficient to provide confidence in the results of the analysis. Although most path models require a minimum of 200 to 300 cases, the answer to this question depends upon the number of parameters being estimated within the model. For more complicated path models, Klem (1995) suggested having at least five to ten observations per estimated parameter. A parameter is defined as an element estimated within the path analysis. Thus, each straight arrow in a path diagram is counted as a parameter. Each double headed arrow and arrow for estimate of error in prediction also count as a parameters to be estimated. Thus, in our case study example we have a total of seven parameters to be estimated. Figure 2 displays four direct effects, one relationship between the two exogenous variables, and two residuals or errors in prediction. The four direct effects are the path coefficients between SAT and GRE, undergraduate GPA and GRE, undergraduate GPA and graduate GPA and GRE and graduate GPA. The relationship between SAT and undergraduate GPA will be estimated. Finally two residual path coefficients will measure the residual or error in prediction for both GRE and graduate GPA. Thus, the minimum sample size that we would need for this analysis ($n = 70 - 200$) is well within the size of our case study sample ($n = 539$). In addition, we find only a moderate relationship between SAT and undergraduate GPA ($r = .483$) and therefore expect limited issues with multicollinearity. Finally, since SAT and undergraduate GPA are measured independently, we expect that the error in the measurement of these variables is unrelated. From an initial assessment, the data for our case study do meet the requirements of the assumptions of path analysis.

As we begin exploring our case study, we must start by reviewing our proposed model (Figure 2). In this model, we have two endogenous variables, GRE and graduate school grade point average, and two exogenous variables, SAT and undergraduate grade point average. In addition, we have the seven parameters to be estimated as described above. So the first step in completing our analysis is to calculate the path coefficients. Path coefficients display the logical link between regression and path analysis, as each path coefficient is the regression coefficient from the appropriate regression analysis. A bivariate or multiple regression analysis is calculated for each endogenous variable in the model. Within each regression analysis the predictor variables are the exogenous variables and the dependent or criterion variable is the relevant endogenous variable. In fact, when a model has an endogenous variable that is predicted from one exogenous variable and bivariate regression is used to calculate an unstandardized path coefficient; the path coefficient will be equal to the correlation coefficient (r). For our case study two multiple regression analyses would be computed. One analysis would specify GRE as the dependent variable and SAT and undergraduate GPA as the predictor variables.

The second would specify graduate GPA as the dependent variable and undergraduate GPA and GRE as the predictor variables. These path coefficients represent the direct effects in the model. The path coefficients are partial regression coefficients that measure the extent of effect of one variable on another in the path model controlling for other prior variables. Similar to regression, these coefficients can be calculated as either standardized or unstandardized coefficients. Path coefficients are most often displayed in standardized format, so that the magnitude of the relationship across different effects can be compared. In path analysis, we distinguish between two different types of standardized coefficients: gamma and beta. Gamma coefficients are those relationships that exist between exogeneous and endogeneous variables, while beta coefficients are those relationships that exist between two endogeneous variables. So in our model we have three gamma (η) coefficients and one beta (β) coefficient.

So when should a researcher present standardized coefficients as opposed to unstandardized coefficients or vice versa? Standardized coefficients are most often displayed when the researcher wishes to describe the relative importance of the predictor variables and the advantage of unstandardized coefficients is that the researcher can describe the impact on the dependent variable of a one-unit change in the predictor variable. Thus, standardized coefficients are often presented on path diagrams and unstandardized coefficients are often presented in summary tables.

While the path coefficients can logically be calculated by running multiple regression analyses, we will use AMOS (Arbuckle, 2003) statistical software to calculate our path analysis case study. AMOS has several advantages over using regression when computing path analysis, the least of which is the ability to run one procedure to calculate the path coefficients for the model. Another advantage to AMOS over regression for calculating path analysis is the ability to test the fit of the model. Figure 3 displays the standardized path coefficients on our path model and Table 1 contains the text output from AMOS with both the standardized and unstandardized coefficients.

Now that we have explored the direct effects, we will explore the indirect effects. The indirect effects are those influences that a variable has on an endogeneous variable that are mediated through other variables in the model. In our model, both SAT and undergraduate GPA have indirect effects on graduate GPA that are mediated through GRE. The paths involved in an indirect effect are sometimes referred to as compound paths as they involve chains of straight arrows, where the path flows in the direction of the arrows. To calculate the indirect effects first locate all of the indirect routes or compound paths and then multiply the path coefficients found on each segment of the indirect route. If only one indirect route exists, as is valid in our case study, then the indirect effect is the product of the path coefficients. If more than one compound path exists then the indirect effect for that variable is equal to the

Figure 3
Case Study: Output Path Diagram

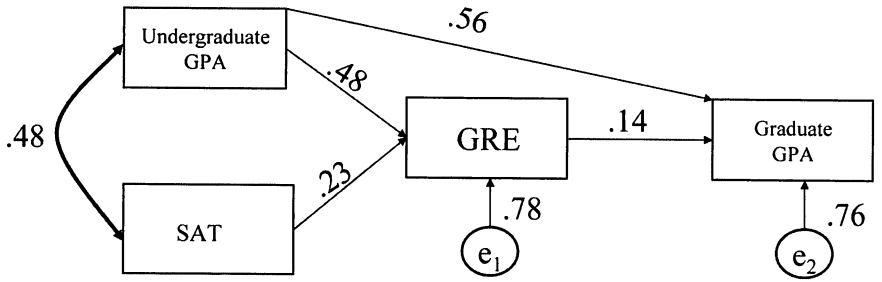


Table 1
AMOS Text Output: Parameter Estimates

Regression Weights: (Group number 1 - Default model)

			Estimate	S.E.	C.R.	P	Label
GRE	<---	Verbal	.274	.047	5.862	***	
GRE	<---	e1	78.847	2.404	32.802	***	
GRE	<---	COLLEGEG	145.936	11.587	12.594	***	
ggpa	<---	E2	.262	.008	32.802	***	
ggpa	<---	COLLEGEG	.574	.042	13.722	***	
ggpa	<---	GRE	.000	.000	3.446	***	

Standardized Regression Weights: (Group number 1 - Default model)

			Estimate
GRE	<---	Verbal	.225
GRE	<---	e1	.781
GRE	<---	COLLEGEG	.484
ggpa	<---	E2	.759
ggpa	<---	COLLEGEG	.558
ggpa	<---	GRE	.140

sum of the products for each path. For our case study, the indirect effect of SAT on graduate GPA is .032 (.225 x .140) and the indirect effect of undergraduate GPA on graduate GPA is .068 (.484 x .140). Remember, a variable in a path model may have only direct effect(s) on one or more variables in the model or may only have indirect effect(s), or may have both. The sum of all direct and indirect effects is sometimes referred to as the total effect or effect coefficient. The total effect for SAT on graduate GPA in our model is .032 or the indirect effect that we just calculated for SAT to graduate GPA, because SAT has no direct effect on graduate GPA. The effect coefficient for undergraduate GPA to graduate GPA is .626 (.558 + .068), which is the sum

of the direct effect of undergraduate GPA to graduate GPA plus the indirect effect that we just calculated. Notice that the total effect for undergraduate GPA has one direct effect and one indirect effect included in the coefficient, while the effect coefficient for SAT has only one component, the one indirect effect. Table 2 displays the text output from AMOS with all the standardized direct, indirect and total effects for both endogenous variables, GRE and graduate GPA.

Table 2
AMOS Text Output: Standardized Effects

Standardized Direct Effects (Group number 1 - Default model)

	COLLEGEG	Verbal	GRE
GRE	.484	.225	.000
ggpa	.558	.000	.140

Standardized Indirect Effects (Group number 1 - Default model)

	COLLEGEG	Verbal	GRE
GRE	.000	.000	.000
ggpa	.068	.032	.000

Standardized Total Effects (Group number 1 - Default model)

	COLLEGEG	Verbal	GRE
GRE	.484	.225	.000
ggpa	.626	.032	.140

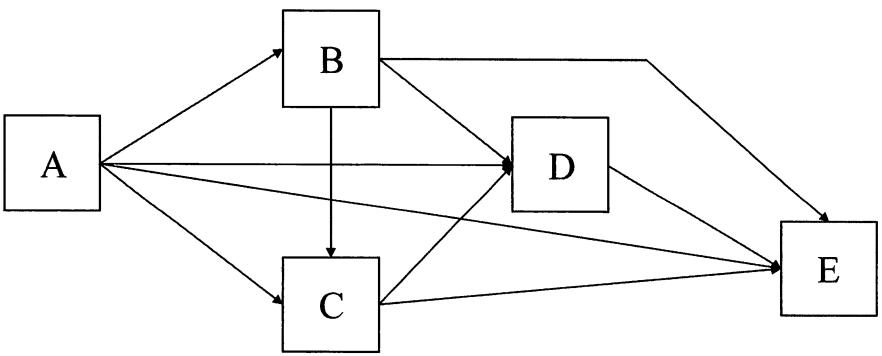
Before we discuss implied coefficients and the fit of the model, we should describe the amount of variance of the endogenous variables that is explained by the predictor variables. On path diagrams with standardized coefficients, the correlation between the error term and the endogenous variable is displayed. We can then square the correlation coefficient to determine the amount of unexplained variance. To determine the amount of explained variance subtract the unexplained from one. For our case study, the amount of variance in Graduate GPA that is explained in our model is .42 or ($1 - .76^2$) and for GRE .39 ($1 - .78^2$). Certainly, our model explains a substantial portion of the variance of each endogenous variable.

An implied correlation, as the name indicates, is the coefficient implied by the model! A coefficient implied by the model consists of four components: direct effect, the sum of the indirect effects, the sum of spurious effects, and the sum of unanalyzed effects. Now we must explore spurious and unanalyzed effects. An unanalyzed effect is an effect that involves the correlation between exogenous variables. Remember, exogenous variables are those variables that the model makes no attempt to predict or explain and a double-headed or curved arrow denotes the relationships between exogenous variables. In our path model, the implied correlation between undergraduate GPA and GRE contains an unanalyzed effect, which is the effect of undergraduate GPA on GRE that flows through SAT. This effect, which is calculated as the

product of two coefficients (.225 x .483), is unanalyzed because it involves a correlation (SAT and undergraduate GPA) for which order is not specified. Spurious relationships occur when variables have common cause. A path that goes against the direction of the arrows on the path diagram characterizes a spurious effect. When a path from one variable to another contains any common cause, the effect is said to be spurious rather than unanalyzed. The implied correlation between any two variables in a model can be calculated by computing each of the individual effects (direct, indirect, spurious, and unanalyzed) and then summing those effects. Analysis of the difference between the implied correlation and the observed correlation will play a large part of the test of the goodness of fit of the model. Before we review the goodness of fit tests we should discuss various types of models.

We must first make the distinction between fully recursive or saturated models and recursive models. In a fully recursive model, each variable has a direct effect on every other variable found below it on the casual chain. Figure 4 displays a fully recursive model. In this model each variable introduced has direct effect on every variable that follows the variable in the model. Thus, the first variable in the model (A) has a direct effect on every subsequent variable in the model. In a model that is recursive, but not fully recursive, one or more of the variables will not have a direct effect on subsequent variables in the sequence of the variables in the model. This distinction is important to note, because a model that is fully recursive will always fit the data perfectly. Thus, the researcher cannot test the fit of a fully recursive model. When first creating models, researchers tend to create fully recursive models due to the fact that every variable is assumed to have some logical connection to every other variable; yet meaningful models are derived from theory and are parsimonious in nature.

Figure 4
Fully Recursive Model



Another important distinction in types of models is between recursive and non-recursive models. A recursive model is one in which all of the effects

are unidirectional and no reciprocal causation exists among variables. A reciprocal causation exists between two variables when variable A is assumed to have a direct effect on variable B and variable B is assumed to have a direct effect on variable A. Thus, two arrows would be found between the variables. In recursive models, the error terms in the model are assumed to be uncorrelated with each other, while in non-recursive models, the error terms must be assumed to be correlated. In this chapter, we are only dealing with recursive models. Again, while non-recursive models may appear to be attractive to the researcher, it is difficult to estimate their parameters and the discussion of these models is beyond the introductory level of this text.

In general terms, the fit of the model is determined by comparing the observed correlation matrix to the implied correlation matrix for the model. Many fit statistics exist and the debate over the correct interpretation of fit statistics in the literature is prolific (Hu & Bentler, 1999; Steiger, 1990; Wheaton, 1987). Within the scope of this text, we will introduce the concept of fit statistics and describe three of the more basic fit statistics: Chi-Square, Normed Fit Index (NFI), and Root Mean Square Error of Approximation (RMSEA). For a more complete description of fit statistics refer to (Hu & Bentler, 1999; Steiger, 1990). The chi-square goodness of fit statistics compares a fitted variance-covariance matrix to the observed variance-covariance matrix on an element-by-element basis. A fitted variance-covariance matrix is calculated from the specifications of the model and is the variance-covariance matrix that would be found, if the model were correct. By comparing these matrices on an element-by-element basis the chi square test is a test of whether or not these residuals are significantly different from zero. As a result, the residuals should be small and not significantly different from zero to indicate a fit of the data to the model. Thus, unlike other statistics where significance is a desired finding, one would want the chi-square to be non significant. The fit statistics from the AMOS text output for our case study are presented in Table 3. As we examine our first fit statistic from our case study, we can see that the residuals are large and our chi-square is significant, which is not a good sign for supporting the fit of our model. Many researchers dismiss the chi-square fit statistic as one that often does not support the fit of a model due to the fact that with large sample sizes small residuals are often statistically significant. On the other hand, small sample sizes often lack the statistical power to find significance; researchers often use both sides of this argument to refute findings that do not support their model. Thus, it is important to explore other fit statistics to provide support for our model. Another chi-square comparison that is commonly made is to compare the chi-square for one's model to the chi-square for an independence model. An independence model is a model that specifies that all variables are uncorrelated. Therefore, the chi-square for this test is most often significant, but to indicate a fit of data to the model, the chi-square for the baseline model must be less than the chi-square for the independence model. For our case

Table 3
AMOS Text Output: Fit Statistics

Model Fit Summary

CMIN

Model	NPAR	CMIN	DF	P	CMIN/DF
Default model	13	50.544	1	.000	50.544
Saturated model	14	.000	0		
Independence model	8	756.300	6	.000	126.050

Baseline Comparisons

Model	NFI Delta1	RFI rho1	IFI Delta2	TLI rho2	CFI
Default model	.933	.599	.934	.604	.934
Saturated model	1.000		1.000		1.000
Independence model	.000	.000	.000	.000	.000

RMSEA

Model	RMSEA	LO 90	HI 90	PCLOSE
Default model	.303	.236	.377	.000
Independence model	.482	.453	.511	.000

study, we have met this standard; as the chi-square for the baseline model, while significant (X^2 = 50.54, $p < .00$) is less than the chi-square for the independence model (X^2 = 756.30, $p < .00$). The Normed Fit Index is a ratio of this comparison. To calculate this statistic, we would first calculate the ratio of the baseline model chi-square to the independence model chi-square and then subtract that ratio from one. For our model NFI is equal to .933, which was calculated as .933 = [1 – (50.54/ 756.30)]. The range of values for fit indices is between 0 and 1.0. In general, fit indices above .90 indicate a fit of the model to the data; values below .90 indicate that the model can be improved. Thus, the NFI for our case study (*NFI* = .933) supports the fit of our model.

The last fit statistic that we will discuss is the Root Mean Square Error of Approximation (RMSEA). This statistic answers the question "how well would the model, with unknown but optimally chosen parameter values, fit the population covariance matrix if it were available" (Browne & Cudeck, 1993, 137-138). A value of about 0.08 or less for the RMSEA would indicate a reasonable error of approximation and would support the fit of the model by the data. Values greater than 0.1 for the RMSEA do not support model fit. For our case study, the RMSEA (.303) does not support the fit of the model. In general, RMSEA favors more complex models and will tend to favor models with many parameters, which does not bode well for our case study model.

So what do we know from our case study? First, the magnitude of our path coefficients and the logic of those coefficients provide support for our proposed model. In Figure 3, we found that Undergraduate GPA was a strong predictor of both GRE (η = .48) and Graduate GPA (η = .56) and interestingly that SAT was a weaker predictor of GRE and GRE (η = .23) was a much weaker predictor of Graduate GPA (β = .14).

Additionally, the explained variance of both endogenous variables is more than reasonable (R^2_{GRE} = .39; R^2_{GGPA} = .42). These direct effects and the magnitude of the explained variance supports the Dean's assertion that performance in undergraduate programs should be emphasized in the selection of students for graduate programs. A limitation to this model is an inconclusive fit of the model to this data. It is important to note that the "Fit" of a model is not a direct test of the magnitude of coefficients or the amount of variance explained in the endogenous variable and fit does not confirm the correctness of the model. Fit of a model is a measure of the extent to which the data align with data that would be implied by the proposed model. Often, fit statistics contradict one another. As we saw in our case study, NFI provides support for our model and RMSEA and Chi-square do not. Even in those cases where multiple fit statistics support the model, one still must not conclude that the model is the best or only model. The model is simply a model that fits the data; other models might be just as good if not better. These other models have simply not been tested or specified yet. So how does a researcher proceed? With caution! Certainly, this preliminary model is interesting and has implications for the institution, yet I would recommend the researcher further explore other models and consider what other variables might add to the explanation of Graduate GPA. As the researcher considers further models, I would emphasize the importance of using theory and model testing. In path analysis and all forms of structural equation modeling the researcher must be guided by theory not data exploration!

Factor Analysis

The second statistical procedure that we will cover is factor analysis. Again as we explore this statistic, we will review the logical theoretical associations between factor analysis and techniques presented in prior chapters. Next, we will describe this technique in depth through the use of a case study application of factor analysis in institutional research. We will end this section with a summary of the interpretation of our case study and a discussion of the application of this technique in institutional research.

Statistical and Theoretical Background

Factor analysis is used to explore the interrelationships among variables to discern whether or not the variables can be grouped into a smaller set of underlying factors. Three primary applications of factor analysis exist. The first purpose is to explore the data for underlying patterns. Factor analysis explores the interrelationships between the items or variables for the presence of underlying factors. Exploratory factor analysis is used for this application. The second application is for data reduction. Factor analysis can be used to reduce a large number of variables into a smaller and more manageable number of factors. Factor analysis can create factor scores for each subject that represent these higher order variables. Exploratory factor analysis is

also used for this application. The third purpose is to confirm the existence of a pre-existing factor structure. When a given factor structure exists within the data, confirmatory factor analysis can be conducted to support the validity of this factor structure. Within institutional research, exploring the data for underlying patterns and reducing data are the primary applications of factor analysis.

Factor analysis is commonly used with surveys or instruments that have many items, some of which could be logically linked to represent one or more higher order factors or constructs. In institutional research, factor analysis is commonly used with senior or alumni surveys including many items designed to assess the outcome of undergraduate experiences. In exploratory factor analysis, the researcher is exploring the data to determine if the variables can be grouped into a smaller set of underlying factors.

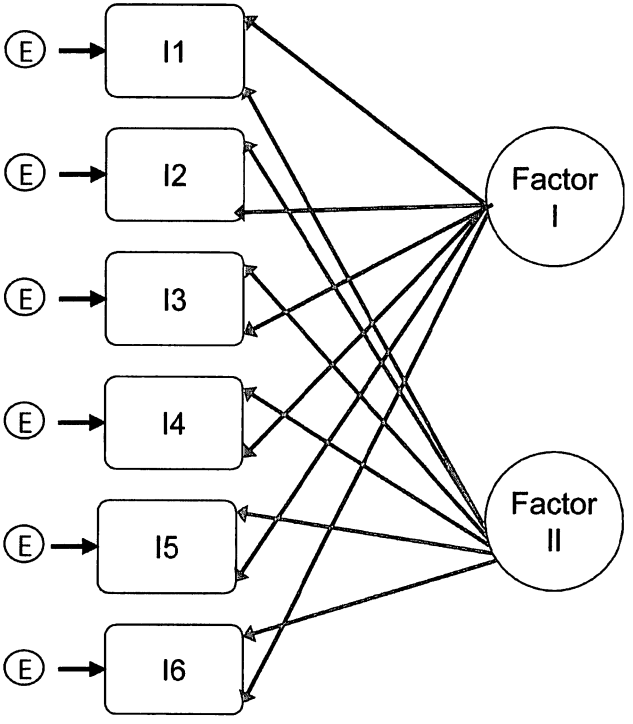
As a result of the exploratory nature of the analysis, all variables are assumed to have a relationship with all factors. Diagrams are also used to display the factor structure. In Figure 5, a theoretical two-factor model, which is being tested using exploratory factor analysis, is displayed. In this model, a total of six items are measured and the model suggests a two-factor structure. Notice that in exploratory factor analysis all items are proposed to have some relationship with both factors, which is noted by the arrow drawn from each factor to each item. In reality, the researcher may have designed the items around the two factor structure and it may logically be that factor one should consist of items one, two, and three, and factor two should consist of items four, five, and six. The model that is tested in exploratory factor analysis does not however, test or confirm the model; it rather explores the data for patterns. Confirmatory factor analysis is used to confirm the existence of a pre-existing factor structure.

Now that we have introduced our exploratory factor model (Figure 5), we have introduced new concepts and terms. In path analysis, we introduced the term "observed variables" as those variables that are measured by the researcher. Remember these variables are displayed as squares in our models. In factor analysis, the observed variables represent the items that are observed or measured. We are exploring the relationships between these items and attempting to group the items into a smaller set of underlying factors. In factor analysis, we now introduce the concept of "latent variables." Latent variables are unobserved variables or hypothetical constructs. These variables are not directly measurable; rather the researcher only has indicators of these measures. These variables are often the more interesting but difficult variables to measure (e.g., leadership, social awareness, or academic achievement). In Figure 5, the latent variables are the two factors. The latent variables are drawn as circles to indicate that they are not directly measured. Notice that the arrows start at the factor and point toward the item. The direction of the arrows is important and is indicative of the fact that the factor or construct is thought to influence the individual's score on the given item;

not the other way around. For example, if the factor is empowerment, and I believe in empowerment, I should rate the items that measure empowerment higher. Because errors in measurement always exist, all items have an error component, which are indicated on our model (Figure 5) as circles that contain an "e."

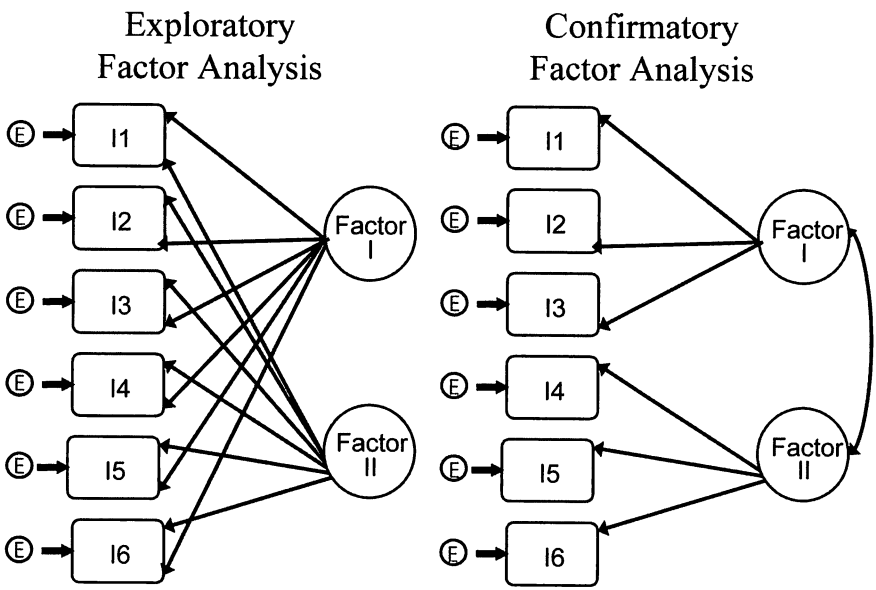
Confirmatory factor analysis is an extension of exploratory factor analysis that meets the final of the three primary applications of factor analysis, which is to confirm the presence or existence of an existing factor structure. Figure 6 compares the exploratory factor analysis that is contained in Figure 5, to a confirmatory factor analysis that would be used to confirm the hypothesized two-factor model. Confirmatory factor analysis differs from exploratory factor analysis in that for a confirmatory analysis, the specific relationships between the items and the factors are confirmed. Certain items are hypothesized to be associated only with given factors; thus not all items have arrows to all factors. Therefore within the confirmatory model found in Figure 6, items one, two, and three are solely associated with factor one, and items four, five, and six are exclusively associated with factor two. The double-headed curved line linking the two factors indicates that a relationship exists between the factors. Again, the circles with the "e" represent the errors in measurement.

Figure 5
Theoretical Two-Factor Exploratory Factor Analysis Model



It is important to note that a researcher should not run an exploratory factor analysis on a data set and then use the same data to confirm the factor structure. Using this procedure one would almost be insured to confirm the structure supported from the exploratory analysis. With large samples,

Figure 6
Comparison of Exploratory to Confirmatory Models



Pedhauzer and Schmelkin (1991) suggest the use of a cross-validation procedure where the researcher randomly splits the sample in half and runs the exploratory analysis on the first sample and the confirmatory analysis on the second sample. In institutional research when we use the same senior (i.e., exit) or alumni survey across multiple years, the researcher may use samples from different years for the exploratory and confirmatory analyses. By using multiple years of data, the researcher can use the results of the exploratory analysis to modify the survey and then use data from the revised instrument to confirm the existing factor structure. Exploratory factor analysis can be completed using many common statistical software packages, such as, SPSS and SAS. Confirmatory factor analysis leads us into structural equation analysis and specialized statistical software, such as, LISREL (Jöreskog & Sörbom, 1999), or AMOS (Arbuckle, 2003).

Case II: Annual Survey of Graduating Students: Outcomes of an Undergraduate Education

To describe the use of exploratory factor analysis our case study will explore the validity of an annual survey of graduating seniors. The Office of Institutional Research at a small private college annually distributes an exit survey to graduating seniors. This instrument consists of a question bank with twenty-seven items that are designed to measure the outcome of an

undergraduate education from this institution. The text of the bank of twenty-seven items may be found in Table 4. From the perspective of the decision makers at the institution, the twenty-seven items are too detailed to individually be of value, thus the researcher wants to determine if the items can be validly organized into a smaller set of underlying factors. While the items were designed to represent a variety of different factors that could be outcomes of an undergraduate education; to date no analysis of the factor structure has been completed. Therefore, the researcher will begin by conducting an exploratory factor analysis. Exploratory factor analysis is the appropriate statistical procedure because the researcher wishes to explore the interrelationships among the items for underlying patterns or factors. In addition, once the factor structure has been established, the researcher wishes to reduce the data by creating factor scores that represent these underlying factors. These two purposes meet the two primary applications of exploratory factor analysis.

Table 4
Senior Survey Items

Write effectively
Communicate well orally
Acquire new skills and knowledge on my own
Think analytically and logically
Formulate creative/original ideas and solutions
Evaluate and choose between alternative courses
Lead and supervise tasks and groups of people
Relate well to people of different races, nations
Function effectively as a member of a team
Use computers for basic tasks (word processing)
Use computers for complex tasks (graphing)
Place current problems in historical prospective
Identify moral and ethical issues
Understand my abilities, my interests, and myself
Function independently without supervision
Gain in-depth knowledge of a field
Plan and execute complex projects
Read or speak a foreign language
Appreciate art, literature, music, and drama
Acquire broad knowledge in the Arts and Sciences
Develop feminist awareness
Develop awareness of social problems
Develop self-esteem /self-confidence
Form close friendships
Establish a course of action to accomplish goals
Synthesize and integrate ideas and information
Understand the role of science and technology

Before exploring our case study further, we should first review the basic statistical assumptions of factor analysis and insure that our data meet the requirements of the statistical procedure. Exploratory factor analysis assumes that the observed variables are a linear combination of some underlying hypothetical or unobservable factors. Some of these factors are assumed to be common to two or more variables and some are assumed to be unique to

each variable. In most exploratory factor analyses, the factors or unobserved variables are assumed to be independent of one another. Some exploratory techniques allow the researcher to account for the relationships that may logically exist between factors; however, these relationships are more commonly accounted for in confirmatory analyses. Finally, all variables in a factor analysis must consist of at least an ordinal scale. Because all of the variables used in the bank of twenty-seven items are ordinal in level of measurement and because the variables are assumed to be a linear combination of some set of underlying factors, the data for our case study meet the requirements of the assumptions of exploratory factor analysis.

It is important to note that nominal data are not appropriate for the type exploratory factor analysis described here. Novice researchers often want to include demographic variables such as gender or ethnicity in factor analysis to account for group differences that may exist; however, this approach is not statistically appropriate. To determine if differences exist between genders on these underlying factors, the researcher would first perform the factor analysis to establish the factor structure and calculate the factor scores. Then the factor scores would be used as the dependent variable in either a t-test or ANOVA design to determine group differences.

The first step in completing an exploratory factor analysis is to measure the interrelationships among the items. This step leads to extraction methods or procedures that determine the appropriate number of factors. When initially determining the appropriate number of factors, one factor is identified for each variable or item. Obviously the researcher expects that the number of useful factors will be substantially less than the total number of items. If, however, no relationship exists between the variables, each variable would make its own unique factor.

Multiple different statistical procedures exist by which the number of appropriate number of factors can be identified. By default SPSS uses the principal-components extraction method. This principal-components method is simpler and was considered by earlier researchers to be the appropriate method of extraction for exploratory factor analysis. Statisticians now advocate the use of other extraction methods due to a flaw in the approach that principal-components utilizes for extraction. Let us briefly explore this issue.

In the principal-components analysis, an inter item correlation coefficient matrix is analyzed to explore the interrelationships between the items and determine if the items can be grouped together to represent a smaller set of underlying factors. The correlation (R) matrix represents the relationships between all items and is a complete matrix with 1.0 on the principal diagonal of the matrix. The 1.0 indicates the perfect relationship the variable has with itself. The upper and lower elements of the matrix contain the shared relationship between each pair of variables and because the relationships are mutually shared the upper and lower elements are mirror images. Table 5 contains a hypothetical correlation matrix representing the relationship

Table 5
Hypothetical Correlation Matrix

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
Item 1	1.000	.678	.780	.240	.189	.065
Item 2	.678	1.000	.858	.380	.345	.188
Item 3	.780	.858	1.000	.189	.243	.058
Item 4	.240	.380	.189	1.000	.789	.834
Item 5	.189	.345	.243	.789	1.000	.657
Item 6	.065	.188	.058	.834	.657	1.000

between six items from a scale. This R matrix reports strong positive relationships among items one, two, and three, as well as strong positive relationships among items four, five and six. In addition, items one, two, and three have weak relationships with items four, five, and six. This pattern would be the first indication that a two-factor model might be appropriate for this data.

The problem with the principal-components extraction method is that correlation matrix has that perfect relationship (r = 1.0) on the principal diagonal. The problem exists because these values (r= 1.0) are used to set the initial communalities. A communality is the extent to which an item correlates with all other items. In principal-components extraction method when the initial communalities are set to 1.0, then all of the variability of each item is accounted for in the analysis. Of course, the flaw is that some of the variability in each item is explained within the factor structure and some is not. Statisticians have indicated that assuming all of the variability of the items whether explained or unique can be accounted for in the analysis is flawed and definitely should not be used in an exploratory factor model.

As a result, principal-axis factoring is suggested as the appropriate method for factor extraction using exploratory factor analysis. In principal-axis factor extraction, the amount of variability each item shares with all other items is determined and this value is inserted into the R matrix replacing the 1.0 on the diagonal of the matrix. As a result, principal-axis factoring is only analyzing common factor variability; removing the uniqueness or unexplained variability from the model. Thus, when reviewing initial communalities using principal-axis factoring, as the values of the communalities decrease more unexplained variability or uniqueness exists within that item. As a result, lower communalities indicate that the item does not add to the proposed factor structure. In layperson's terms, this is a problem item not common to the proposed factor structure, but is more unique or outside the factor structure.

Table 6 contains the SPSS output for the initial communalities from our case study data using both principal-components and principal-axis extraction procedures. Using the principal-components extraction methods, one can clearly see that all of the initial communalities are set to 1.0, indicating the flawed assumption that the model explains all of the variance of each item. Reviewing the initial communalities using principal axis factor extraction, one will notice all of the items have initial communalities substantially less than 1.0. The items with the lowest communalities include: read or speak a foreign language, use computers for basic tasks, and use computers for complex tasks. These items do not add to the proposed factor structure. If the researcher wanted to shorten the bank of twenty-seven items, these items could be considered for deletion or if the items were unclear, the researcher may wish to consider rewording the items. On the other hand, just because the communalities are low, does not mean that the item should automatically be removed. Because these items appear to be clear and logical, we will retain them in our analysis. Later on in our case we will consider other criteria for revising or removing items.

While principal-components analysis is not suggested as a factor extraction method and principal-axis factoring is suggested for all the reasons cited above, other extraction methods are also available. Two of the other more commonly used extraction methods are the generalized least-squares method and the maximum likelihood method. The generalized least squares factor extraction method minimizes the sum of the squared differences between the observed and reproduced correlation matrices. A reproduced correlation matrix shows the predicted pattern of relationships between the items when the factor analysis solution is assumed to be correct. Thus, when the factor structure is supported by the data, the reproduced correlations will be close to the observed values. In the generalized least squares extraction method, correlations are weighted by the inverse of their uniqueness, so that variables with high uniqueness are given less weight than those with low uniqueness. The maximum-likelihood factor extraction method produces parameter estimates most likely to have produced the observed correlation matrix if the sample is from a multivariate normal distribution. The correlations are also weighted by the inverse of the uniqueness of the variables, and an iterative algorithm is employed. The main advantage of these extraction methods over principal-axis factoring is that these extraction methods can be used to produce a goodness-of-fit test for the analysis. A goodness-of-fit test is used to test the significance of the model. This test can be calculated when these extraction procedures are used because each of these procedures creates parameter estimates that represent the proposed model (e.g., reproduced correlation matrix), which the observed data can be tested against (i.e., goodness of fit). Thus, the goodness-of-fit test indicates whether or not the proposed factor analysis model fit the data.

Table 6
SPSS Output: Case Study Communalities

Communalities

	Initial	Extraction
Write effectively	1.000	.453
Communicate well orally	1.000	.437
Acquire new skills and knowledge on my own	1.000	.488
Think analytically and logically	1.000	.555
Formulate creative / original ideas and solutions	1.000	.541
Evaluate and choose between alternative courses	1.000	.538
Lead and supervise tasks and groups of people	1.000	.712
Relate well to people of different races, nations	1.000	.480
Function effectively as a member of a team	1.000	.671
Use computers for basic tasks (word processing)	1.000	.545
Use computers for complex tasks (graphing)	1.000	.667
Place current problems in historical prospective	1.000	.539
Identify moral and ethical issues	1.000	.564
Understand myself, my abilities, interests	1.000	.609
Function independently without supervision	1.000	.548
Gain in-depth knowlegde of a field	1.000	.345
Plan and execute complex projects	1.000	.484
Read or speak a foreign language	1.000	.634
Appreciate art, literature, music, drama	1.000	.602
Acquire broad knowledge in the Arts and Sciences	1.000	.483
Develop feminist awarenenss	1.000	.597
Develop awareness of social problems	1.000	.685
Develop self-esteem /self-confidence	1.000	.626
Form close friendships	1.000	.592
Establish a course of action to accomplish goals	1.000	.572
Synthesize and integrate ideas and information	1.000	.523
Understand the role of science and technology	1.000	.591

Extraction Method: Principal Component Analysis.

Communalities

	Initial	Extraction
Write effectively	.366	.361
Communicate well orally	.377	.369
Acquire new skills and knowledge on my own	.412	.401
Think analytically and logically	.396	.459
Formulate creative / original ideas and solutions	.462	.489
Evaluate and choose between alternative courses	.443	.467
Lead and supervise tasks and groups of people	.412	.623
Relate well to people of different races, nations	.347	.367
Function effectively as a member of a team	.408	.486
Use computers for basic tasks (word processing)	.256	.266
Use computers for complex tasks (graphing)	.254	.383
Place current problems in historical prospective	.331	.352
Identify moral and ethical issues	.440	.464
Understand myself, my abilities, interests	.461	.535
Function independently without supervision	.418	.446
Gain in-depth knowlegde of a field	.298	.248
Plan and execute complex projects	.408	.409
Read or speak a foreign language	.153	.181
Appreciate art, literature, music, drama	.320	.547
Acquire broad knowledge in the Arts and Sciences	.309	.342
Develop feminist awarenenss	.353	.369
Develop awareness of social problems	.496	.643
Develop self-esteem /self-confidence	.535	.593
Form close friendships	.329	.398
Establish a course of action to accomplish goals	.536	.537
Synthesize and integrate ideas and information	.522	.493
Understand the role of science and technology	.435	.525

Extraction Method: Principal Axis Factoring.

As we continue with our case study we have employed principal-axis factoring and the extraction procedure has extracted twenty-seven factors, one for each of the twenty-seven items. Now we need to determine if the number factors that explain a substantial amount of the total variance is significantly less than twenty-seven. To complete the extraction process, eigenvalues are calculated and interpreted for each factor. Eigenvalues represent the amount of variance in the data that is explained by the factor with which it is associated. Eigenvalues have common characteristics. First, in principal-axis factor extraction, the factors are extracted in order by the amount of variance they explain. Therefore, the first factor will have the highest eigenvalue, the second the next highest, through to the last factor and eigenvalue, which will explain the least amount of variance. Second, the first few factors generally explain the majority of the variance with the last few explaining only a very small proportion of variance. Table 7 contains the SPSS output for our eigenvalues using the principal-axis factor extraction method. In this table, the above describe characteristics are visible. The first panel lists the initial eigenvalues for all twenty-seven items. The first factor most definitely explains the largest amount of the variability (30.78%). From the initial extraction, the first six factors explain over 55% of the total variance with the remaining twenty-one factors explain the remaining 45% of the variance.

Determining the optimal number of factors to extract is not a straightforward task because the decision is ultimately subjective. Several criteria exist for determining the number of factors to be extracted, but these are just empirical guidelines rather than an exact quantitative solution. One such guideline is the eigen-one or Kaiser-Guttman rule. This rule instructs the researcher to keep only those factors whose eigenvalues are greater than 1.0 and discard the rest. The rationale for choosing the value of 1.0 is that a factor must account for variance at least as large as the variance of a single standardized variable. Remember, standardized variables have a mean of zero and a standard deviation and variance of 1.0. While this rule is generally accepted, some statisticians have indicated that the eigen-one rule is still somewhat arbitrary and that the researcher should be driven by theory when determining the appropriate number of factors to extract. Another way to determine the number of useful factors is visually with a scree plot. Eigenvalues are plotted on the Y or vertical axis of the scree plot and the factors are plotted on the horizontal or X-axis. The scree plot can be used to help determine the optimal number of factors or components to retain in the solution, as the scree plot should form the intersection of two lines. Factors on the initial steep line of the plot should be retained and factors on the scree, which is the gradual trailing line, should be eliminated. Figure 7 contains the scree plot for our case study data. Although the intersection of the two lines is not clear, the scree is identifiable. Because our analysis is not driven by any specific theory or model and our scree plot is not clear, we will apply the eigen-one rule to our case study data.

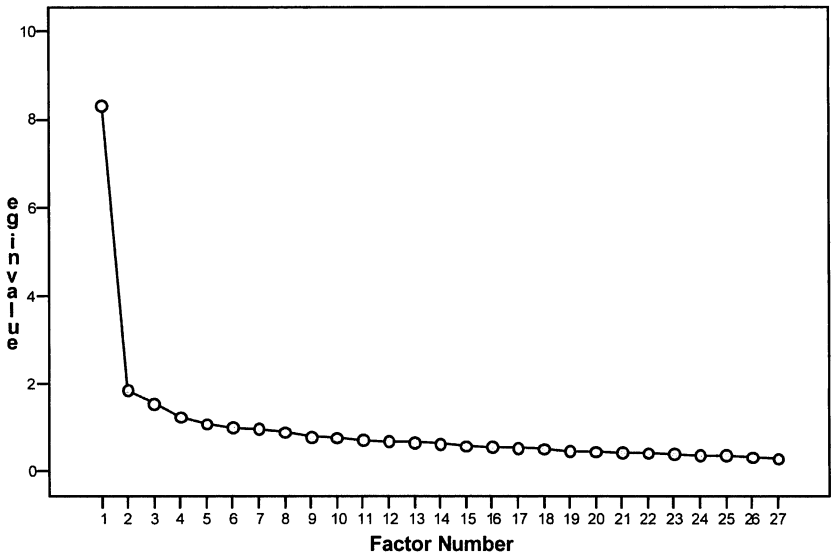
Table 7
SPSS Output: Case Study Eigenvalues

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Factor	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	8.311	30.783	30.783	7.774	28.793	28.793	3.880	14.369	14.369
2	1.859	6.885	37.667	1.315	4.871	33.664	2.043	7.568	21.936
3	1.556	5.762	43.430	.998	3.695	37.359	1.901	7.040	28.976
4	1.252	4.638	48.067	.648	2.399	39.758	1.501	5.558	34.535
5	1.095	4.054	52.121	.537	1.989	41.746	1.303	4.825	39.359
6	1.007	3.731	55.852	.483	1.788	43.535	1.127	4.176	43.535
7	.977	3.619	59.471						
8	.905	3.352	62.823						
9	.795	2.946	65.769						
10	.769	2.850	68.618						
11	.718	2.661	71.279						
12	.686	2.540	73.819						
13	.668	2.475	76.293						
14	.632	2.342	78.636						
15	.580	2.147	80.782						
16	.567	2.098	82.881						
17	.537	1.987	84.868						
18	.518	1.920	86.788						
19	.468	1.734	88.521						
20	.453	1.679	90.201						
21	.437	1.618	91.819						
22	.425	1.575	93.393						
23	.403	1.491	94.884						
24	.376	1.392	96.276						
25	.372	1.376	97.652						
26	.333	1.233	98.885						
27	.301	1.115	100.000						

Extraction Method: Principal Axis Factoring.

Figure 7
Scree Plot



Applying the eigen-one rule leaves us with a total of six extracted factors. Given the extracted number of factors, the eigenvalues are then recalculated based on the extracted factors only. The second panel in Table 7 displays these values. In this panel, we have only the six extracted factors. If we had used principal components extraction, the values for the initial and extracted eigenvalues would be the same. Because, we used principal-axis extraction these values will not be identical to the initial eigenvalues. Extraction methods other than principal-components generally produce eigenvalues smaller than the initial values, due to errors in measurement.

After the number of meaningful factors has been determined, the researcher must begin to interpret the factor structure. To interpret the factor structure and place meaning to the factors or constructs that have been established, a factor matrix is calculated. Table 8 contains the initial factor matrix for our case study data. This matrix identifies the relationships between the variables and the factors. In general, this matrix reads like a correlation matrix and the elements of the matrix (i.e., factor loadings) are similar in structure and value to correlation coefficients. As a general rule, factor loadings of .40 or greater (either positive or negative) indicate that the item is associated with that factor. Before, interpreting this initial factor matrix, we must be aware of some general problems with this initial factor matrix. Three problems exist with this initial factor structure. As stated earlier, the first factor explains

Table 8
SPSS Output: Initial Factor Matrix

Factor Matrix [a]

	Factor					
	1	2	3	4	5	6
Write effectively	.523	-.033	-.274	-.048	.094	.020
Communicate well orally	.581	.042	-.076	-.112	-.091	.051
Acquire new skills and knowledge on my own	.558	.174	-.232	-.074	-.015	.024
Think analytically and logically	.541	.193	-.279	-.129	.164	-.086
Formulate creative / original ideas and solutions	.632	.117	-.202	-.152	.038	.099
Evaluate and choose between alternative courses	.631	.098	.046	-.163	.067	.162
Lead and supervise tasks and groups of people	.520	.106	.453	-.151	.001	.338
Relate well to people of different races, nations	.517	-.146	.277	.008	.027	-.034
Function effectively as a member of a team	.524	.065	.410	-.039	.062	.183
Use computers for basic tasks (word processing)	.379	.135	.208	.171	-.004	-.177
Use computers for complex tasks (graphing)	.206	.409	.238	.293	.136	-.112
Place current problems in historical prospective	.479	-.244	-.114	.031	.156	.158
Identify moral and ethical issues	.593	-.246	.025	.074	.215	.010
Understand myself, my abilities, interests	.634	-.229	.031	-.166	-.189	-.128
Function independently without supervision	.620	.044	.048	-.041	-.209	-.110
Gain in-depth knowlegde of a field	.434	.169	-.172	.020	.028	-.013
Plan and execute complex projects	.570	.256	-.133	.017	-.015	.036
Read or speack a foreign language	.230	-.097	-.104	.236	-.202	.107
Appreciate art, literature, music, drama	.462	-.277	-.179	.296	-.276	.246
Acquire broad knowledge in the Arts and Sciences	.444	.084	-.104	.332	-.101	.076
Develop feminist awarenenss	.380	-.428	-.019	.123	.158	-.036
Develop awareness of social problems	.574	-.439	.075	.147	.281	-.121
Develop self-esteem /self-confidence	.716	-.155	.063	-.075	-.134	-.170
Form close friendships	.473	-.150	.205	-.092	-.232	-.219
Establish a course of action to accomplish goals	.717	.083	.027	-.076	-.047	-.087
Synthesize and integrate ideas and information	.661	.149	-.133	-.049	.083	-.081
Understand the role of science and technology	.509	.431	.060	.269	.017	-.060

Extraction Method: Principal Axis Factoring.

a. 6 factors extracted. 17 iterations required.

the most amount of variance and as a result most of the variables will have at least some relationship with this first factor. Thus, this factor becomes very generalized and difficult to interpret. Second, many factors may be bipolar. A bipolar factor is one in which both significant positive and negative loadings exist. A negative loading is like a negative correlation coefficient. Often negative loadings or relationships may be found that are due to the actual values in the data. Bipolar factors can, however, create negative loadings that cannot be interpreted logically from the data. Finally, because of the first factor being a general factor, many variables may load on more than one factor, creating double factor loadings. While this complexity is not a problem statistically, the question of whether it adds needlessly to the complexity of the factor structure arises.

Table 9 contains the initial factor matrix that has been sorted by size and only factor loadings of .40 or greater are printed. Using these options makes identification of the problems with the initial matrix easier. In this table, examples of the problems described above are visible. The first factor is most definitely a general factor. All items except four have factor loadings above .40 on this factor. The only items that do not load on this first factor are: use computers for basic tasks, develop a feminist awareness, use computers for complex tasks, and read or speak a foreign language. Three items have double factor loadings (high loadings on two factors). Those three items are: function effectively as a team, lead and supervise tasks and groups of people, and understand the role of science and technology. In addition the item, develop awareness of social problems, revealed a bipolar double factor loading. Additionally, the item, develop a feminist awareness, has a negative factor loading on factor two. To address these three issues statisticians suggest that the initial factor matrix not be analyzed and that a rotation procedure be completed prior to interpreting the factor structure. Three common procedures exist for rotation: orthogonal, oblique, and varimax. Each method varies in how the rotation is accomplished. Let us briefly review the concept of rotation and these rotation procedures before returning to the analysis of our case study.

Rotation is a process that is used to simplify the interpretation of a factor analysis. The concepts and principles of rotation can most clearly be explained by reviewing a factor plot. A factor plot of two hypothetical factors is shown in Figure 8. The axes lines represent the values of the factor loadings for two factors. The two factors are placed at right angles to one another, because in the first stage of exploratory factor analysis the factors are assumed to be unrelated. The Xs represent the variables and are plotted where the factor loadings for the two factors intersect. In this figure, many items have high factor loadings on both factors. Negative factor loadings appear for some variables on factor 1. Factor analysis starts with the original axes from the extraction process and then applies a mathematical rotation that simplifies the relationships between the items and the factors. How that

Table 9
SPSS Output: Initial Factor Matrix – Sorted and Suppressed

Factor Matrix[a]

	Factor					
	1	2	3	4	5	6
Establish a course of action to accomplish goals	.717					
Develop self-esteem /self-confidence	.716					
Synthesize and integrate ideas and information	.661					
Understand myself, my abilities, interests	.634					
Formulate creative / original ideas and solutions	.632					
Evaluate and choose between alternative courses	.631					
Function independently without supervision	.620					
Identify moral and ethical issues	.593					
Communicate well orally	.581					
Develop awareness of social problems	.574	-.439				
Plan and execute complex projects	.570					
Acquire new skills and knowledge on my own	.558					
Think analytically and logically	.541					
Function effectively as a member of a team	.524		.410			
Write effectively	.523					
Lead and supervise tasks and groups of people	.520		.453			
Relate well to people of different races, nations	.517					
Understand the role of science and technology	.509	.431				
Place current problems in historical prospective	.479					
Form close friendships	.473					
Appreciate art, literature, music, drama	.462					
Acquire broad knowledge in the Arts and Sciences	.444					
Gain in-depth knowlegde of a field	.434					
Use computers for basic tasks (word processing)						
Develop feminist awarenenss		-.428				
Use computers for complex tasks (graphing)		.409				
Read or speack a foreign language						

Extraction Method: Principal Axis Factoring.

a. 6 factors extracted. 17 iterations required.

Figure 8
Initial Relationship Between Hypothetical Two-Factor Structure and Items

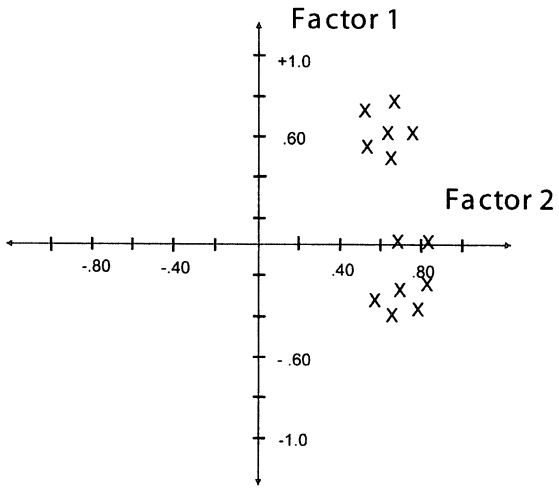
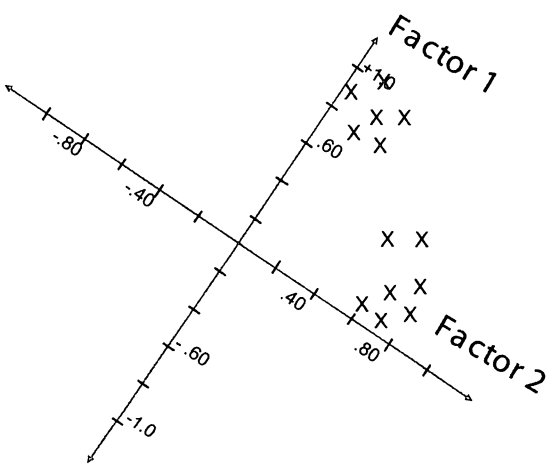


Figure 9
Orthogonal Rotation



mathematical rotation is performed varies based upon the rotation procedure.

Orthogonal rotation is easiest of the three methods to describe conceptually; however, this method is the most limited in terms of its application. Orthogonal rotation has a restriction that states that factors may only be rotated in such a manner that the factors are kept at right angles to each other. This restriction follows the assumption that no association exists between the factors. For many applications within institutional research the factors are logically correlated, thus the limited application of this procedure.

Orthogonal rotation can best be described graphically. As orthogonal rotation is completed, the factors are now rotated or shifted to improve the relationships between the variables and each factor while maintaining the right angle restriction. Graphically the orthogonal rotation procedure is illustrated in Figure 9. The rotation procedure increased the identification of the uniqueness of each factor and reduced the number of items with negative and double factor loadings.

Using oblique rotation, factors are rotated without maintaining the right angle restrictions. This procedure allows the researcher to account for the

relationships that may logically exist between the factors. The closer the rotation angle is to zero, the higher the correlation between the factors. In fact, if the rotation angle were zero, the two factors would have merged, as the lines would have fallen on top of one another on our factor plot. As the rotation angle approaches 90^0, the relationship between the factors nears zero and approximates the orthogonal rotation method. Again, oblique rotation can best be described visually. Figure 10 displays an oblique rotation between these same two factors and items. In this figure notice that the right angle restriction is no longer held; however, only a slight relationship is expected between the two factors, as the factors are still at about an 80^0 angle. In exploratory factor analysis, little is known about the relationship between the factors, which can limit the usefulness of this procedure.

Our third rotation procedure is varimax rotation. Although varimax is an orthogonal rotation, which also assumes no relationship between the factors, the procedure is designed to maximize the amount variance uniquely accounted for by each factor. The varimax method minimizes the number of variables that have high loadings on each factor and as a result simplifies the interpretation of the factors. This procedure is a commonly applied technique. After all, shouldn't explaining the maximum amount of variability in the scores by each factor and simplifying the factor structure be a priority? For our case study we will use varimax rotation.

Figure 10
Oblique Rotation

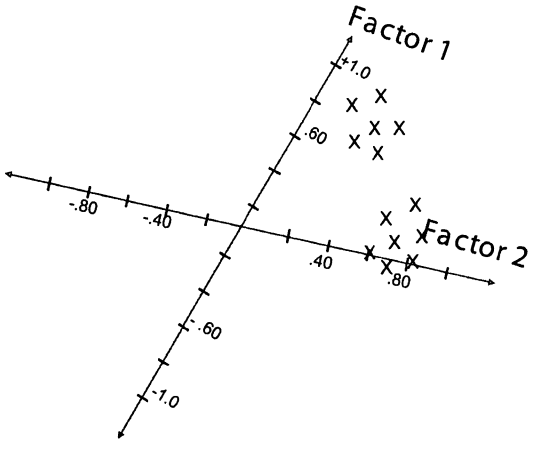


Table 10 contains the rotated factor component matrix for our case data. The table has again been sorted by size and had factor loadings less than .40 suppressed. In contrast to Table 9 with the initial component matrix, one can see that for our data the majority of the problems with the initial matrix have been addressed. When interpreting a factor matrix, the researcher needs to try to find the common thread or theme that items loading (i.e., have factor loadings greater than .40) on that factor have in common. The thread or theme is then interpreted as the name of the factor or latent variable.

Before naming the factors many researchers support applying the principles of Thurstone's simple structure. Thurstone's simple structure is a set of general guidelines that help the researcher interpret a rotated factor

Table 10
SPSS Output: Rotated Factor Matrix – Sorted and Suppressed

Rotated Factor Matrix [a]

	Factor					
	1	2	3	4	5	6
Think analytically and logically	.643					
Formulate creative / original ideas and solutions	.625					
Synthesize and integrate ideas and information	.587					
Acquire new skills and knowledge on my own	.581					
Plan and execute complex projects	.537					
Write effectively	.512					
Establish a course of action to accomplish goals	.496					
Evaluate and choose between alternative courses	.485					
Communicate well orally	.460					
Gain in-depth knowlegde of a field	.442					
Develop awareness of social problems		.732				
Develop feminist awarenenss		.560				
Identify moral and ethical issues		.542				
Place current problems in historical prospective		.433				
Form close friendships			.569			
Understand myself, my abilities, interests			.547			
Develop self-esteem /self-confidence			.543			
Function independently without supervision			.455			
Relate well to people of different races, nations						
Lead and supervise tasks and groups of people				.724		
Function effectively as a member of a team				.567		
Use computers for complex tasks (graphing)					.601	
Understand the role of science and technology					.571	
Use computers for basic tasks (word processing)					.402	
Appreciate art, literature, music, drama						.650
Acquire broad knowledge in the Arts and Sciences						.413
Read or speak a foreign language						.402

Extraction Method: Principal Axis Factoring.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 8 iterations.

matrix. Thurstone's guidelines tell the researcher to select items that relate strongly to the proposed factor (i.e., factor loadings of .40 or above), and to delete or drop items that are double loaded (i.e., .40 or above on more than one factor). Next, items that are unique or do not load on any factor (i.e., all factor loadings are below .40) are deleted. Finally, items that load high on a factor that was not the proposed factor for that item are deleted. As one can see by these guidelines, the term simple structure implies that items should be related to only one factor. Some statisticians acknowledge that factor complexity is inevitable and that many factor structures may have items that should logically be related to more than one factor. As a result, some researchers will maintain items with double factor loadings as long as the items would logically belong to both factors. In cases when greater than .05 percent difference exists between the two factor loadings, the item can be considered as primarily belonging to the factor with the higher factor loading.

Our rotated factor matrix (Table 10) does not reveal any items with double loadings, nor are there any items that do not load on any factor. While we did not have a predetermined factor structure, logical relationships do exist between these items, and a review of the rotated factor matrix does not reveal any items that should be deleted for loading on an illogical factor. Our data support a clean 6-factor structure.

If the researcher had difficulty interpreting the rotated factor matrix and problems were revealed, then reviewing the extracted communalities may be helpful in determining the fate of some items. Remember, the communalities represent the extent to which an item correlates with all other items. The extracted communalities can be calculated from either the initial or rotated factor matrix. The extracted communality for any one item is calculated by squaring the rotated factor loadings and then summing those squared values across all factors. Table 11 displays the calculations that would be done to the rotated factor matrix to create the extracted communalities for our case study. Remember, the factor loadings represent the relationship shared between the item and the factor, thus the square of the relationship represents the amount of explained variability. This concept parallels the relationship between a correlation coefficient and the coefficient of determination or R-squared. Therefore, the lower the communality the more unexplained variability or uniqueness exists within that item. So when an item is presenting problems within the factor structure and the communality for that item is low, the researcher should most definitely consider revising or removing the item from the survey.

Now we must try to name these factors or latent variables. The naming of the factors is the responsibility of the researcher. Naming of the factors is the most important and difficult stage in the interpretation of factor analysis. When naming the factors consider items that have higher factor loadings as being more representative of the factor than items with lower factor loadings. In some cases, the researcher has some predetermined factor structure

Table 11
Calculation of Eigenvalues and Communalities

	1	1 Square	2	2 Square	3	3 Square	4	4 Square	5	5 Square	6	6 Square	Communalities
Write effectively	.5228	.2733	-.0335	.0011	-.2744	.0753	-.0476	.0023	.0938	.0088	.0200	.0004	.3612
Communicate well orally	.5814	.3380	.0421	.0018	-.0763	.0058	-.1123	.0126	-.0910	.0083	.0512	.0026	.3691
Acquire new skills and knowledge on my own	.5579	.3113	.1744	.0304	-.2315	.0536	-.0736	.0054	-.0147	.0002	.0235	.0006	.4015
Think analytically and logically	.5415	.2932	.1927	.0371	-.2793	.0780	-.1290	.0166	.1639	.0269	-.0864	.0075	.4593
Formulate creative / original ideas and solutions	.6325	.4000	.1175	.0138	-.2018	.0407	-.1519	.0231	.0382	.0015	.0987	.0098	.4889
Evaluate and choose between alternative courses	.6307	.3978	.0982	.0096	.0465	.0022	-.1634	.0267	.0671	.0045	.1621	.0263	.4670
Lead and supervise tasks and groups of people	.5196	.2699	.1063	.0113	.4529	.2051	-.1507	.0227	.0012	.0000	.3380	.1142	.6233
Relate well to people of different races, nations	.5171	.2674	-.1464	.0214	.2766	.0765	.0084	.0001	.0271	.0007	-.0343	.0012	.3673
Function effectively as a member of a team	.5244	.2750	.0646	.0042	.4101	.1682	-.0392	.0015	.0620	.0038	.1826	.0334	.4861
Use computers for basic tasks (word processing)	.3788	.1435	.1350	.0182	.2078	.0432	.1711	.0293	-.0037	.0000	-.1772	.0314	.2656
Use computers for complex tasks (graphing)	.2057	.0423	.4090	.1672	.2383	.0568	.2928	.0857	.1362	.0186	-.1119	.0125	.3832
Place current problems in historical prospective	.4787	.2291	-.2436	.0594	-.1143	.0131	.0308	.0010	.1562	.0244	.1582	.0250	.3519
Identify moral and ethical issues	.5927	.3513	-.2461	.0605	.0254	.0006	.0741	.0055	.2150	.0462	.0100	.0001	.4643
Understand myself, my abilities, interests	.6342	.4023	-.2287	.0523	.0305	.0009	-.1656	.0274	-.1891	.0358	-.1279	.0164	.5351
Function independently without supervision	.6201	.3846	.0436	.0019	.0482	.0023	-.0414	.0017	-.2092	.0438	-.1101	.0121	.4464
Gain in-depth knowlegde of a field	.4336	.1880	.1694	.0287	-.1718	.0295	.0200	.0004	.0277	.0008	-.0131	.0002	.2476
Plan and execute complex projects	.5695	.3244	.2557	.0654	-.1332	.0177	.0168	.0003	-.0145	.0002	.0359	.0013	.4093
Read or speak a foreign language	.2299	.0529	-.0971	.0094	-.1040	.0108	.2365	.0559	-.2020	.0408	.1069	.0114	.1813
Appreciate art, literature, music, drama	.4621	.2135	-.2769	.0767	-.1792	.0321	.2963	.0878	-.2758	.0761	.2462	.0606	.5468
Acquire broad knowledge in the Arts and Sciences	.4443	.1974	.0836	.0070	-.1043	.0109	.3322	.1104	-.1012	.0102	.0758	.0057	.3416
Develop feminist awarenenss	.3796	.1441	-.4281	.1833	-.0193	.0004	.1229	.0151	.1584	.0251	-.0355	.0013	.3692
Develop awareness of social problems	.5736	.3290	-.4395	.1931	.0748	.0056	.1468	.0216	.2813	.0792	-.1206	.0145	.6430
Develop self-esteem /self-confidence	.7160	.5127	-.1546	.0239	.0628	.0039	-.0748	.0056	-.1341	.0180	-.1700	.0289	.5930
Form close friendships	.4726	.2234	-.1496	.0224	.2055	.0422	-.0921	.0085	-.2318	.0537	-.2191	.0480	.3982
Establish a course of action to accomplish goals	.7169	.5140	.0826	.0068	.0267	.0007	-.0755	.0057	-.0470	.0022	-.0866	.0075	.5369
Synthesize and integrate ideas and information	.6611	.4371	.1493	.0223	-.1332	.0178	-.0485	.0024	.0827	.0068	-.0807	.0065	.4928
Understand the role of science and technology	.5087	.2588	.4309	.1857	.0601	.0036	.2694	.0726	.0171	.0003	-.0599	.0036	.5246
Eigenvalues -- Rotated		7.774151		1.315078		0.997648		0.647768		0.536904		0.482887	

used in this phase. Remember, when a proposed factor structure does exist, the researcher would eventually want to complete a confirmatory factor analysis. Later in this case study, we will briefly introduce confirmatory factor analysis. Because our case study does not have a predetermined factor structure, we are looking for the common thread, theme, or construct shared by items with large factor loadings for each of our extracted factors. When naming factors, do not do so in isolation. Be sure that it is done in the context of the research and organization. Get input from respected colleagues and relevant campus groups as factors are named. The most important element in naming the factors is that when the factor and the items that make up each factor is reported, individuals must agree that the items fit under this theme or construct. Name recognition is what the researcher is seeking.

Table 12
Factors and Associated
Senior Survey Items

Intellectual
- Think analytically and logically
- Formulate creative / original ideas and solutions
- Synthesize and integrate ideas and information
- Acquire new skills and knowledge on my own
- Plan and execute complex projects
- Write effectively
- Establish a course of action to accomplish goals
- Evaluate and choose between alternative courses
- Communicate well orally
- Gain in-depth knowledge of a field

Moral
- Develop awareness of social problems
- Develop feminist awareness
- Identify moral and ethical issues
- Place current problems in historical prospective

Self-Development ~ Self-Awareness
- Form close friendships
- Understand my abilities, my interests, and myself
- Develop self-esteem /self-confidence
- Function independently without supervision

Leadership
- Lead and supervise tasks and groups of people
- Function effectively as a member of a team

Technology
- Use computers for complex tasks (graphing)
- Understand the role of science and technology
- Use computers for basic tasks (word processing)

Humanities
- Appreciate art, literature, music, and drama
- Acquire broad knowledge in the Arts and Sciences
- Read or speak a foreign language

After discussion and consultation with various constituencies at our institution, these factors were named Intellectual, Moral, Self-Development ~ Self-Awareness, Leadership, Technology, and Humanities, respectively. Table 12 lists the factor name and the items that comprise each factor. It is important to respect the naming process. At a different institution, the names of these factors might be

different, yet the content would remain the same. Remember, name recognition and acceptance of the logic or face validity of the factor structure is the ultimate objective.

As we complete our factor analysis, we should go back and take one final look at Table 7, which contains the eigenvalues for our case study. The third panel of the table contains the eigenvalues or summary of explained variability for the final rotated factor structure. The third panel is titled *rotated sum of squares loadings*. This label is appropriate because the eigenvalues are calculated by squaring the rotated factor loadings (Table 11) and then summing those squared values across all items. Again, remember that the factor loadings represent the relationship shared between the item and the factor and that the square of the relationship represents the amount of explained variability. When reviewing these values, one will notice that the variance accounted for by each of the factors in the rotated structure does not equal the variance accounted for by the extracted matrix. This change is due to the rotation procedures that were applied. After all, the goal of all rotation procedures is to make the relationships between the factors and the items clearer. While the distribution of explained variability is adjusted across factors, the cumulative or total amount of explained variability will remain the same from the extracted to the rotated factor structure. In our case study, the total amount of the variability of the scores across the twenty-seven items that is explained by both the extracted and rotated factor structures equals 43.5%.

Once the factor analysis is complete and the factors have been named, the researcher may want to use these factors in reporting data from the survey and explore for potential differences in these factors across various subgroups in the sample. To do so the researcher should create factor scores. Factor scores quantify individual scores for each participant on each of the factors or latent variables. Several methods exist for creating factor scores. Two common methods for creating factor scores that are available in most statistical packages are regression method and Anderson Rubin method. The regression method creates predicted scores for the factors that have a mean of zero and a variance equal to the squared multiple correlations between the estimated factor scores and the true factor values. These scores may be correlated even when factors are orthogonal. If the researcher is interested in how each item contributes to the calculation of the factor scores you can review the factor score coefficient matrix. The factor score coefficient matrix shows the values used to compute factor scores for each case. For each case, the factor score can be computed by multiplying standardized variable values (z-scores) by the factor score coefficients and then summing these values across the factors. Each item is weighted by the coefficient to represent its contribution to the factor. For principal-component models, these procedures are followed and exact component scores are calculated. For other extraction methods, standard regression scores cannot be computed. Table 13 contains the factor structure coefficients for our case study. The coefficients are used

Table 13
SPSS Output: Factor Structure Coefficients

Factor Score Coefficient Matrix

	Factor					
	1	2	3	4	5	6
Write effectively	.175	.068	-.101	-.057	-.091	-.005
Communicate well orally	.104	-.073	.049	.043	-.096	.041
Acquire new skills and knowledge on my own	.202	-.071	-.045	-.036	-.051	.019
Think analytically and logically	.298	.033	-.096	-.122	-.018	-.199
Formulate creative / original ideas and solutions	.262	-.040	-.109	.060	-.154	-.027
Evaluate and choose between alternative courses	.142	-.011	-.097	.220	-.106	-.069
Lead and supervise tasks and groups of people	-.094	-.092	-.110	.743	-.086	-.006
Relate well to people of different races, nations	-.102	.092	.105	.086	.067	-.044
Function effectively as a member of a team	-.097	.030	-.043	.368	.079	-.044
Use computers for basic tasks (word processing)	-.079	.011	.106	-.062	.246	-.016
Use computers for complex tasks (graphing)	-.072	.019	-.065	-.035	.463	-.040
Place current problems in historical prospective	.062	.178	-.184	.073	-.110	.061
Identify moral and ethical issues	.003	.287	-.120	.025	.028	-.046
Understand myself, my abilities, interests	-.003	-.049	.402	-.053	-.180	-.008
Function independently without supervision	.004	-.134	.273	-.050	.030	.065
Gain in-depth knowlegde of a field	.113	-.020	-.049	-.050	.036	.014
Plan and execute complex projects	.152	-.081	-.062	-.009	.061	.056
Read or speack a foreign language	-.047	-.036	-.007	-.020	.011	.273
Appreciate art, literature, music, drama	-.098	-.032	-.082	.037	-.123	.755
Acquire broad knowledge in the Arts and Sciences	-.020	-.019	-.084	-.051	.161	.310
Develop feminist awarenenss	-.059	.273	-.048	-.049	-.016	.023
Develop awareness of social problems	-.133	.652	-.076	-.119	.144	-.153
Develop self-esteem /self-confidence	-.023	-.009	.418	-.108	-.023	-.025
Form close friendships	-.111	-.074	.410	-.062	.011	-.026
Establish a course of action to accomplish goals	.093	-.052	.173	-.021	.047	-.065
Synthesize and integrate ideas and information	.197	.020	-.014	-.085	.059	-.109
Understand the role of science and technology	.024	-.081	-.070	-.075	.483	.112

Extraction Method: Principal Axis Factoring.
Rotation Method: Varimax with Kaiser Normalization.
Factor Scores Method: Anderson-Rubin.

to compute the factor scores for each case by multiplying variable values by the factor score coefficients.

The problem in interpreting these regression factor scores is that they are not on a standardized scale and even when an orthogonal rotation is used the factor scores may be correlated. The Anderson Rubin method ensures orthogonality of the estimated factor scores and the scores produced are standardized with a mean of zero, a standard deviation of one. Thus, this method makes interpretation of these scores easier as the factor scores can be interpreted as z-scores, ranging from approximately -3.0 to +3.0. Because of the complexity of dealing with data on a standardized scale, some researchers shy away from using factor scores and would rather create scale scores based upon the mean of the raw scores for the items associated with each factor. Two problems exist in using this approach. First, this approach should not be used if items used in the factor analysis are based upon different response scales. The second problem is that by using the mean of the items, the researcher is assuming that each item contributes equally to the construct or factor. By reviewing the factor loadings we know that this assumption is flawed. Therefore, factor scores have less error in estimating the construct or latent variable and are preferable. The researcher may now use these factor scores to determine if differences exist in these six factors between any subgroups (e.g., gender) in the population.

So what do we know from our case study? At this point we have determined that our data support a six-factor model. The factor structure revealed was relatively clean and free of double loadings and interpretation problems. We can now use this factor structure to discuss the findings from our survey data. We may organize descriptive statistics using the factor structure and/or we could further explore our data using the factor scores to determine if differences exist across various sub-groups on the factor scores. What should we do with our factor analysis as we move forward with this survey? The next logical step would be to use the survey again the following year and subsequently run confirmatory factor analysis to confirm the factor structure and test the fit of the model to this new set of data. Our next section of this chapter will introduce confirmatory factor analysis and structural equation modeling.

Introduction to Structural Equation Modeling

Structural equation modeling is a statistical approach to testing hypotheses about the relationships between observed or measured variables and latent variables (Hoyle, 1995). As we begin to explore structural equation modeling, let us again begin by reviewing the theoretical and logical associations that exist between structural equation modeling and techniques presented in prior chapters and earlier sections of this chapter. Next, we will describe this technique in depth by briefly extending our prior case study application of factor analysis. We will end this section with a summary of the

interpretation of our case study and a discussion of the implications for structural equation modeling in institutional research.

Statistical and Theoretical Background

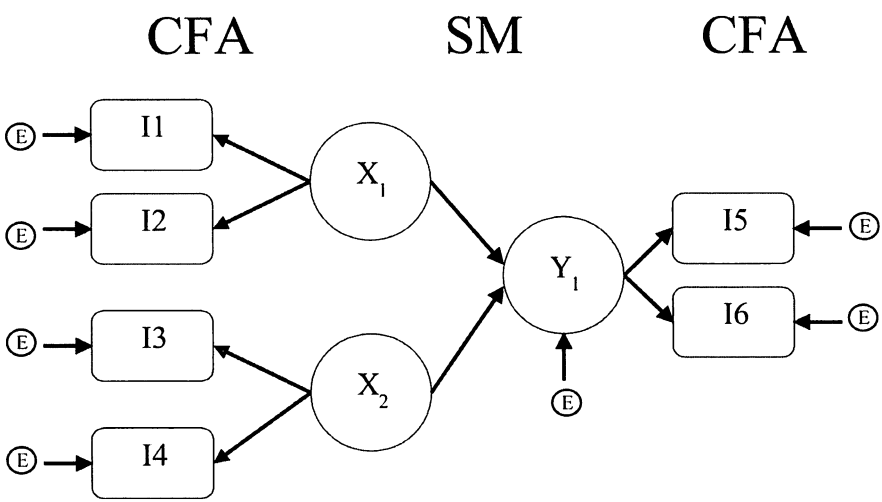
Three types of structural equation models (SEM) are most applicable to institutional research. The first type of structural equation model is path analysis. Some statisticians do not consider path analysis to be a form of structural equation modeling, because path analysis consists of only observed variables. In contrast, others, (myself included), do believe that path analysis should be considered within the frame of structural equation models for two important reasons. Most importantly, path analysis uses the same underlying principles of model specification and model fitting as other structural equation models. Secondly, path analysis is an important part of the historical development of SEM and most researchers use SEM software to analyze path models.

The second type of structural equation model is Confirmatory Factor Analysis. As discussed in our prior section, confirmatory factor analysis is an extension of exploratory factor analysis used to confirm a hypothesized factor structure. Remember, confirmatory factor analysis differs from exploratory factor analysis in that for a confirmatory analysis, the specific relationships between the items and the factors are confirmed. Certain items are hypothesized to be associated only with given factors; thus not all items have arrows to all factors (Figure 6). As a result, confirmatory factor analysis truly fits our definition of structural equation modeling as the researcher is testing a hypothesis about the relationships between observed or measured variables and latent variables.

The third type of structural equation model is a structural regression model. Structural regression models, also commonly referred to as structural equation models, can be viewed as a logical extension of confirmatory factor analysis. While confirmatory factor analysis confirms the hypothesized relationships between observed and latent variables, structural regression models allow the researcher to additionally explore the relationships between the latent variables. Figure 11 logically extends our confirmatory factor analysis found in Figure 6 by proposing a third factor and suggesting the relationships between the three factors. By exploring this figure we can see the difference between confirmatory factor analysis and structural regression models. Our model now consists of two components: measurement model(s) and structural model. In this analysis, we have two measurement models and one structural model. The two measurement models represent the two confirmatory factor analyses that would establish the measurement of our latent variables and the structural model identifies the relationships between the latent variables.

Whether running path analysis, confirmatory factor analysis, or full structural regression models, we see both similarities and differences between SEM and other standard statistical procedures such as correlation, multiple

Figure 11
Hypothetical Structural Equation Model



regression, and ANOVA. Before we explore the essence of SEM through a brief example of confirmatory factor analysis, let us summarize some similarities and differences. Hopefully through our prior discussion of path analysis, the logical link that exists between correlation, regression, and path analysis were evident. In fact, some statisticians (Hoyle, 1995; Raykov & Marcoulides, 2000; Schumacker & Lomax, 1996) have indicated that standard linear models are special instances of the general structural equation model. One similarity between these techniques is, that like these standard statistical techniques, statistical tests associated with SEM are valid only if the assumptions regarding the observed data are met. For SEM, the most common assumptions are independence of observations and multivariate normality of observed data. Another important similarity between SEM and standard statistical techniques is that neither approach offers a statistical test of causality. SEM in its early years was sometimes referred to as causal modeling because these techniques enjoy some advantage over more restrictive standard models. In this regard, none of these techniques including SEM can be used to imply causality, as this is a condition only established through logic, strong theory, or methodological strategies.

SEM differs from standard statistical approaches like regression and ANOVA in three important ways. First, unlike these techniques, SEM requires the researcher to formally specify the model to be tested. Hoyle (1995) best summarized this distinction, "unlike ANOVA, which, as typically used, evaluates main effect and interaction hypotheses by default and multiple regression analysis, which permits specification only of direct effects on a

single outcome, SEM offers no default model specification and places relative few limits on what types of relations can be specified" (p. 14). A second and commonly cited advantage of SEM over standard statistical techniques is the ability of SEM to test the relationships between latent variables. Of the standard statistical procedures, only exploratory factor analysis can be used to estimate latent variables and none of these techniques can be used to test the relationships among latent variables. The third distinction between SEM and standard statistical techniques is really a weakness of SEM; the ambiguity and debate over the statistical tests of the fit of structural equation models compared to the relatively straightforward and commonly accepted inferential tests associated with techniques such as ANOVA and regression.

When testing any structural equation model using any of the specialized structural equation software (e.g., AMOS; Arbuckle, 2003; LISREL, Jöreskog & Sörbom, 1999), the researcher must complete several steps. First, the researcher must specify the model to be tested. Model specification is a key part of structural equation modeling. The researcher must carefully think about his or her data and hypothesize the relationships associated with each variable in the model. I cannot stress enough that theory and logic are paramount in SEM. As discussed earlier in our discussion of path analysis and exploratory factor analysis, path diagrams are used to display the model specifications.

The next phase in the analysis is determining the model parameters to be estimated. Model parameters reflect those aspects of a model that are unknown to the researcher at the beginning of the analysis, yet are needed to test the model. Parameter is, of course, a generic term that refers to some characteristic of a whole population (e.g., the population mean). In SEM, the model parameters are the unknown aspects of the hypothesized or specified model that are estimated by the distributions of the observed variables in the model. In other words, they are the elements that will be estimated from the sample covariance or correlation matrix using the specialized computer software. These parameters are estimated in such a way that they can be tested against a model that would emulate the specified model. The parameters may be estimated using one of several estimation procedures, which we will discuss in our case study example. The goal of these estimation procedures or numerical routines is basically to minimize the fit of the data to the specified model. These numerical routines then proceed in a consecutive or iterative manner to select values for model parameters in such a way that at each step or iteration the distance between the population value and the sample value is reduced.

Finally, the process ends when no further improvement in the fit function can be achieved, or no decrease occurs in the difference between the population and sample values. At that point in time, final parameter estimates are produced and various fit statistics are calculated. The task of the researcher

is then to interpret those parameters and fit statistics to determine if the specified model fit the data. Remember, even in those cases where model parameters are logical and supportive of the model and fit statistics support the model, the researcher may not conclude that this model is the "best or only model." Rather, the researcher should conclude that this is a logical model supported by the data, but other models that have yet to be theorized or tested could exist and in fact may be superior to the tested model. Now, let us specifically review these steps using our prior case study example.

Case III: Confirming the Factor Structure from the Annual Survey of Graduating Students

For illustrative purposes, we are going to continue with our second case study example, but we are going to reduce the number of factors and items in our model. When first working with SEM models smaller and more parsimonious models are easier for the researcher. As we will see in this case study and as is the case in the real world, more complicated and less parsimonious models are necessary to describe human behavior. Therefore, the main role of this case study is for illustrative purposes. Also, we must again point out that we would not run the confirmatory factor analysis on the same data sets that we ran the exploratory factor analysis. To run the analysis in this manner would be redundant and would almost insure a fit to the model. As a result, we would need a new sample from a subsequent data collection.

For the purposes of this case study, we will explore a subset of our Senior Survey with three factors. The factors and the items selected are illustrated on Figure 12. For this model, the three factors include, intellectual, leadership and technology. The factors are associated with five, three and three items, respectively. Before we discuss the model parameters to be estimated, some important basic path diagram guidelines need to be reviewed. Latent variables are our hypothetical constructs or factors. In the path diagram, circles denote these latent variables. Observed or measured variables are denoted in the path diagram as rectangles and are the items from our senior survey. Remember, the lines or paths that are drawn between the factors and items start at the factor because it is the hypothetical construct that is thought to influence or drive the response of the individual to the item, not the other way around. The path diagram also displays the errors in measurement for each of the items and each of the latent variables. Now we have specified our confirmatory factor analysis.

The next step in the analysis is to determine the model parameters to be estimated. Almost all of the specialized SEM software (e.g., AMOS; Arbuckle, 2003; LISREL, Jöreskog & Sörbom, 1999) will take the specified model, determine the parameters to be estimated and present a summary of the model parameters. The summary of our model parameters for our case study may be found in Table 14. It is important that the researcher check the model parameters to determine that the model has been correctly specified.

Bentler (1995) summarized six rules that can be used to determine the parameters of the model. Rule 1 states that all the variances of the exogeneous variables are model parameters. In addition, Rule 1 also states that the variances of the error terms associated with the latent variables are model parameters. Although in my model, the error terms have been standardized or set to have a mean of zero and a standard deviation of one. As a result for our case study rule 1 states that we have eleven variances that are model parameters to be estimated and three fixed variances associated with our factors, therefore, we have a total of fourteen variances in our parameter summary. Rule 2 states that all covariances between exogeneous variables are model parameters. Remember, these covariances would be indicated by double-headed curved arrows between our observed variables. In confirmatory factor analysis we do not estimate the covariances between the items, so we do not have any covariances as model parameters. Rule 3 states that all factor loadings are parameters to be estimated, thus we have eleven factor loadings to be calculated, one for each observed variable. Rule 4 states that

Figure 12
Path Diagram for Confirmatory Factor Analysis

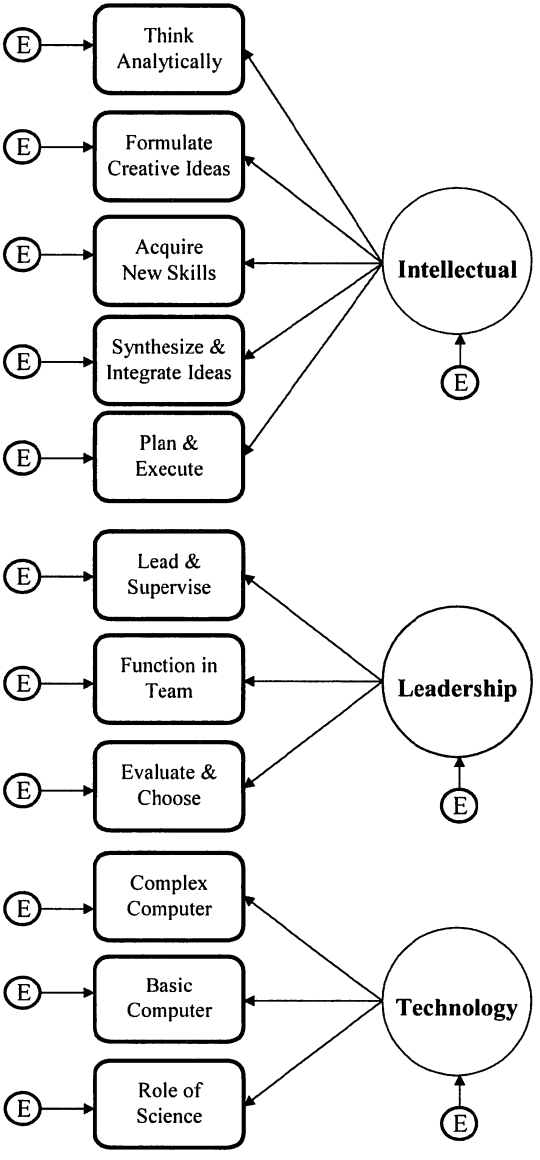


Table 14
AMOS Output: Model Summary

Parameter summary (Group number 1)

	Weights	Covariances	Variances	Means	Intercepts	Total
Fixed	14	0	3	0	0	17
Labeled	0	0	0	0	0	0
Unlabeled	11	0	11	0	11	33
Total	25	0	14	0	11	50

all regression coefficients are parameters to be estimated. This rule includes both the weight and the intercept for all latent variables including error terms. Although it is important to note that the error terms have been standardized with a mean of zero and a standard deviation of one; the weights are listed under fixed and the constants for these terms are not model parameters. Also it is important to note that rule 4 can be thought of as a special case of rule 3, as factor loadings are really a coefficient or weight between the item and the factor.

Rule 5 states that variances and covariance between the endogenous variables are not model parameters. This rule is due to the fact that these estimates are in fact determined by other model parameters. In any case, our confirmatory factor analysis does not specify any covariances between the endogenous variables as we have no curved double headed arrows between the latent variables. Rule 6 states that for each latent variable included in the model the metric of its scale must be specified. This rule is commonly met by scaling the latent variable to one of the observed variables. When the results of an exploratory factor analysis are available, set the path between the latent variable and the observed variable with the highest factor loading to 1.0. This is another reason why three of our weights or coefficients are fixed. Thus, we end up with fifty model parameters, seventeen of which are fixed and associated with our fourteen error terms (fourteen weights or coefficients and three variances), and thirty-three of which are free to be estimated and associated with our eleven items (eleven weights, eleven intercepts, and eleven variances).

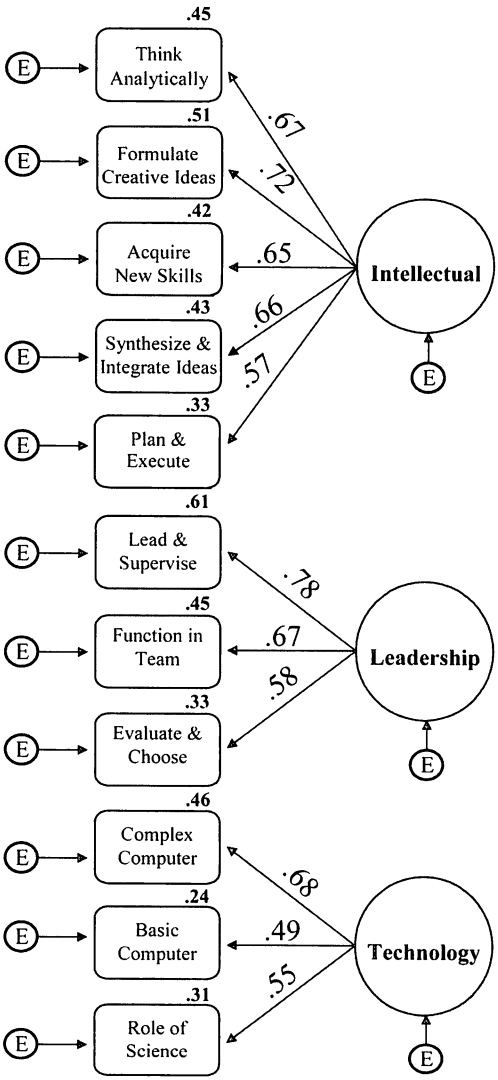
Now that we know we have correctly identified our model we are ready to move to a discussion of the estimation method and model identification. Four main estimation methods are commonly used by specialized SEM software. The four methods are unweighted least squares, maximum likelihood, generalized least squares, and weighted least squares. Each of the procedures is based upon some form of the sum of the squared difference between the corresponding elements of the observed variance-covariance matrix to the reproduced variance-covariance matrix. Recall from our earlier discussion of fit statistics, a reproduced variance-covariance matrix contains the variances and covariances that are emulated or implied by the model. An

unweighted least square estimation method is the unweighted sum of the squared differences between each of the corresponding elements of the sample variance-covariance matrix and the reproduced matrix. The maximum likelihood method or generalized least squares are commonly applied when the data meet the assumption of multivariate normality. The weighted or asymptotically distribution-free estimation method should be used when assumptions of normality are not met, although large sample sizes are required for this application. The maximum likelihood method is more commonly used and has been shown to handle slight variations in the assumption of normality. Maximum likelihood is the procedure commonly used as the default estimation procedure in most SEM software programs. As a result it is the estimation procedure we will use in our case study.

Using the estimation procedure the iterative process is implemented until such time as the process has converged and the final model is identified. Remember, the iterative process ends when no further improvement in the fit function can be achieved or no decrease occurs in the difference between the elements of the sample variance-covariance and the reproduced variance-covariance matrices. Before moving on to discuss the interpretation of the parameter estimates and fit statistics for our case study, it is important to spend a few moments discussing parameter and model identification. A model parameter may be unidentified if not enough empirical information exists to estimate the specific parameter. A model that contains even one unidentified parameter cannot be interpreted, even though some elements of the model may be logical and useful. So what does having an unidentified parameter mean? Having an unidentified parameter implies that no method exists to determine a unique value for the estimate. It is much like having two unknowns in one algebraic equation. Given that SEM models are attempting to estimate on an element-by-element basis the difference between the sample variance-covariance matrix and the reproduced matrix, it is quite possible that with some data and models given parameters may have many solutions. In these cases unidentified parameters exist and the model is unidentified. Sometimes unidentified models are the result of misspecification. Recall the six rules for determining the model parameters. When any one of these rules is violated, an unidentified model will result.

Although it is important to state that following these rules is a necessary, but not sufficient requirement for having an identified model. In other words, a researcher may follow all the model specification rules and still the data do not allow for the identification of a given parameter. Normally, if a proposed model is carefully conceptualized, the chances of unidentified parameters are minimal. Often this message occurs when the researcher has failed to set the scale on a latent variable (i.e., Rule 6) and thus is easily fixed within the SEM software program. When a more serious version of under identification occurs the researcher may need to totally reconsider the model or consider additional data collection. It is important to stress again the importance of theory and logic in SEM.

Figure 13
Path Diagram with
Parameter Estimates



Now that the model from our case study has been identified and the parameters and fit statistics have been determined, we must proceed to analyze the output from our analysis. The parameter estimates are commonly presented on the path diagram. Figure 13 presents the path diagram with the parameter estimates from our case study. From this figure we can see that all of the factor loadings are above .40, indicating a strong association between each of the latent variables and the observed items. Given that we conducted an exploratory factor analysis, we would have expected this finding with our cross validation sample. Another statistic presented on the path diagram is the squared multiple correlation between each latent and observed variable. These values, which are presented on the top of each of the observed variables, are strong, ranging from .24 to .61. To interpret our highest squared multiple correlation between the factor of leadership and the item, lead and supervise tasks and groups of people, we could state that the factor leadership explain approximately 61 percent of the variance of the item. In other words, the error variance of item, lead and supervise tasks and groups of people, is approximately 39.2 percent. Thus, these values also support the logic of our model. So now we must turn to the evaluation of the fit statistics for our model.

Again, the concept of the evaluation of the fit of the model is complex. As mentioned earlier, many fit statistics exist and the debate over the correct interpretation of fit statistics in the literature is prolific (Hu & Bentler, 1999). Given the applied nature of this text, we will only interpret the three basic fit statistics, which were discussed in the context of path analysis: Chi-Square, Normed Fit Index (NFI) and Root Mean Square Error of Approximation (RMSEA). The fit statistics for the confirmatory factor analysis defined in our case study may be found in Table 15. Let's start with the chi-square goodness-of-fit statistic, which compares a fitted variance-covariance matrix to the observed variance-covariance matrix on an element-by-element basis. Remember, the chi square test is a test of whether or not these residuals are significantly different from zero and, therefore, the residuals should be small and not significantly different from zero to indicate a fit of the model to the data. Unfortunately, the chi-square for our model is quite large and significant ($X^2 = 568.52$, $p < .00$). We can again compare our chi-square for our model to an independence model and see that the chi-square for our model is less than the independence model ($X^2 = 1858.89$, $p < .00$), which is a minimal expectation of a reasonable fit. However, the NFI statistic for our case study ($NFI = .697$), which is a ratio of the above describe comparison, does not fall within the desired range (> .90). Thus, we have another fit statistic that does not support the fit of our model.

The RMSEA is our third fit statistic. Remember, a value of about 0.08 or less for the RMSEA would indicate a reasonable error of approximation and would support the fit of the model to the data. Values greater than 0.1 for the RMSEA do not support model fit. We see more bad news for our model, as the RMSEA for our case study is .137. One may ask, why do we present a case study with fit statistics that do not support the model? The answer is that in practice many researchers will find that their initial models have poor fit statistics. Also remember that while our case study comes from an initial exploratory factor analysis, we have simplified and reduced the model for illustrative purposes. In fact, when I run the full model on the data the parameter estimates and fit statistics do improve, although the fit statistics are still contradictory.

So what can we draw from this last case study? Given the abbreviated nature of our model, the practical implications from our case are outweighed by the illustrative values. My hope is that those of you new to Structural Equation Modeling have been presented with an overview of the technique and an example applied within institutional research. One may be wondering outside of confirmatory factor analysis what role does SEM have in institutional research? I truly think that SEM will become more prevalent in research on the outcomes of higher education and institutional research. A growing need exists for us to measure and define the relationships between our observed variables and the more interesting outcome measures that are latent variables (e.g., intellectual growth, leadership, social awareness). Again as we venture

Table 15
AMOS Output: Fit Statistics Confirmatory Factor Analysis

Model Fit Summary

CMIN

Model	NPAR	CMIN	DF	P	CMIN/DF
Default model	33	568.520	44	.000	12.921
Saturated model	77	.000	0		
Independence model	11	1858.895	66	.000	28.165

Baseline Comparisons

Model	NFI Delta1	RFI rho1	IFI Delta2	TLI rho2	CFI
Default model	.694	.541	.711	.561	.707
Saturated model	1.000		1.000		1.000
Independence model	.000	.000	.000	.000	.000

RMSEA

Model	RMSEA	LO 90	HI 90	PCLOSE
Default model	.137	.127	.147	.000
Independence model	.207	.199	.215	.000

into such areas, I must once again stress the need for theory and logic to guide our research and our practices. It is this process that will move our research from confirmatory models into structural regression models, where we will test the relationships between our latent variables. As I look at my confirmatory model, I can think of proposing some relationships (e.g., does intellectual growth influence leadership), but I will leave that discussion to texts at the next level.

References

Arbuckle, J. L. (2003). *AMOS 5.0 Update to the AMOS User's Guide.* Chicago: SmallWaters.

Bentler, P. M. (1995). *EQS structural equation program manual.* Encino, CA: Multivariate Software.

Browne, M., & Cudeck, R. (1993). Alternative ways of assessing mode fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models.* Newbury Park, CA: Sage.

Hoyle, R. H. (1995). The structural equation modeling approach: Basic concepts and fundamental issues. In R. H. Hoyle (Ed.), *Structural equation modeling: concepts, issues and applications.* Newbury Park, CA: Sage

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventioal versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55.

Klem, L. (1995). Path analysis. In L. G. Grimm & P. R. Yarnold (Eds.). *Reading and understanding multivariate statistics* (pp. 65-96). Washington, DC: American Psychological Association.

Jöreskog, K. G., & Sörbom, D. (1999). *LISREL 8.30: User's reference guide.* Chicago: Scientific Software.

Pedhauzer, E. J., & Schmelkin L. (1991). *Measurement, design, and analysis: An integrated approach.* Hilsdale, NJ: Lawrence Erlbaum Associates.

Raykov, T. & Marcoulides, G. (2000). *A first course in structural equation modeling.* Mahwah, NJ: Lawrence Erlbaum Associates.

Schumacker, R. E., & Lomax, R. G. (1996). *A beginners guide to structural equation modeling.* Malwah, NJ: Lawrence Erlbaum Associates.

Steiger, J. H. (1990). Structural model evaluation: An interval estimation approach. *Multivariate Behavioral Research, 25,* 173-180.

Wheaton, B. (1987). Assessment of fit in overidentified models with latent variables. *Sociological Methods and Research, 16,* 118-154.