February 2004

A POLICYMAKER'S PRIMER ON EDUCATION RESEARCH

How To Understand, Evaluate and Use It

A joint effort of Mid-continent Research for Education and Learning and the Education Commission of the States

Understanding Statistics Tutorial

Overview

Statistics refers to methods and rules for organizing and interpreting quantitative observations. The purpose of this tutorial is to explain basic statistical concepts commonly used in education research. The goal is to help readers understand the results reported in quantitative education research.

Descriptive Statistics

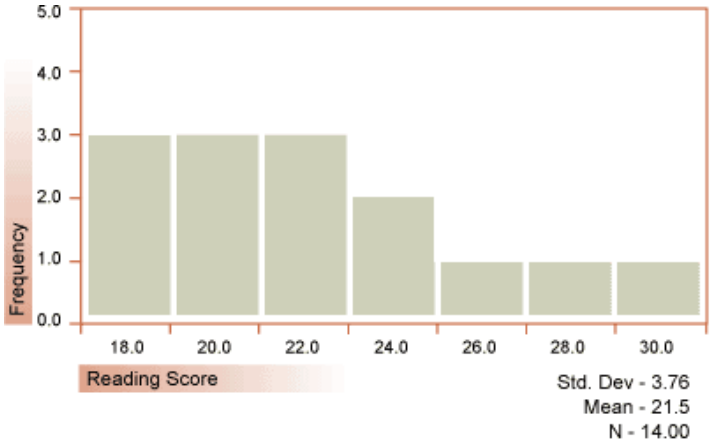
Descriptive statistics are used to describe sets of numbers such as test scores. Researchers organize sets of scores into tables and graphs called frequency distributions.

Example 1:
The following numbers represent students' scores on a reading test: 19,23,17,27,21,20,17,22,19,17,25,21,29,24

A frequency table shows the distribution or number of students who achieved a particular score on the reading test. In Example 1, three students achieved a score of 17.

Reading Score	Frequency	Percent	Percentile
17	3	21.4	21.4
19	2	14.3	35.7
20	1	7.1	42.9
21	2	14.3	57.1
22	1	7.1	64.3
23	1	7.1	71.4
24	1	7.1	78.6
25	1	7.1	85.7
27	1	7.1	92.9
29	1	7.1	100.0
TOTALS	**14**	**100.0**	

A frequency graph also shows the distribution or number of students who achieved a particular score.



The following are the most common statistics used to describe frequency distributions:

N – the number of scores in a population

n – the number of scores in a sample

Percent – the proportion of students in a frequency distribution who had a particular score. In Example 1, 21% of the students achieved a score of 17.

Percentile – The percent of students in a frequency distribution who scored at or below a particular score (also referred to as percentile rank). In Example 1, 79% of the students achieved a score of 24 or lower, so a score of 24 is at the 79th percentile.

Mean – The average score in a frequency distribution. In Example 1, the mean score is 21.5. (Abbreviations for the mean are M if the scores are from a sample of participants and μ if the scores are from a population of participants.)

Median – The score in the middle of frequency distribution, or the score at the 50th percentile. In Example 1, the median score is 21.

Mode – The score that occurs most frequently in the distribution. In Example 1, the mode is 17.

Range – The difference between the highest and lowest score in the distribution. In Example 1, the range is 12.

Standard Deviation – A measure of how much the scores vary from the mean. In the sample, the standard deviation is 3.76, indicating that the average difference between the scores and mean is around 4 points. The higher the standard deviation, the more different the scores are from one another and from the mean. (Abbreviations for the standard deviation are *SD* if the scores are from a sample and Σ if the scores are from a population.)

The mean, median and mode are called measures of central tendency because they identify a single score as typical or representative of all the scores in a frequency distribution.

When a frequency distribution has a high standard deviation, the mean is not a good measure of central tendency as in the following set of scores:

Example 2:
Scores = 1,4,3,4,2,7,18,3,7,2,4,3
Mean = 5
Median = 3.5
Standard Deviation = 4.53

The standard deviation in Example 2 indicates that the average difference between each score and the mean is around 4.5 points. Only one score (18) however is 4.5 or more points different from the mean. In this example, the one extreme score (18) overly influences the mean. The median (3.5) is a better measure of central tendency because extreme scores do not influence the median.

Standard Score – Specifies the location of an original score or raw score within a frequency distribution, based on standard deviation units. Standard scores also are known as *z*-scores and are calculated as follows:

z = (Raw Score – Mean) /Standard Deviation.

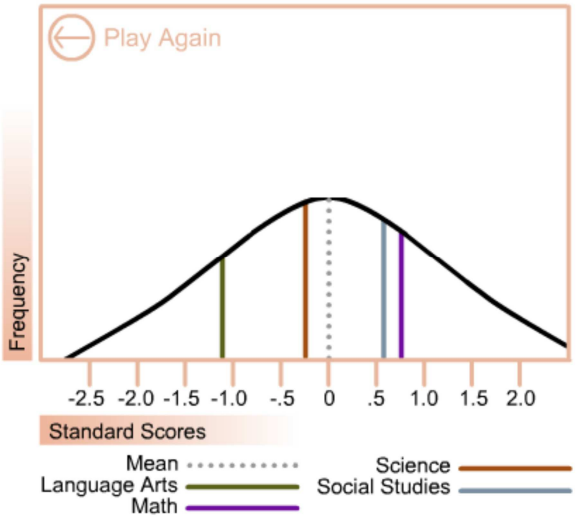
In Example 1, a raw score of 27 has a standard score of +1.46 (27 – 21.5 / 3.76). This indicates that a score of 27 is 1.46 standard deviation units above the mean. A raw score of 19 has a standard score of –.66, indicating that it is .66 standard deviation units below the mean.

Standard scores make it possible to compare scores on different tests that have different means and standard deviations. For example, the following table shows a student's raw scores and standard scores on four different tests.

Subject	Raw Score	Standard Score
Mathematics	31	+ .75
Language Arts	71	– 1.10
Science	42	– .25
Social Studies	42	+ .56

On which test did this student perform best in comparison to the rest of the students in the class? Numerically, the student's

highest score was on the language arts test, but the standard score for language arts indicates that the student performed worst on this test because the score was 1.1 standard deviation units below the mean. The student's best performance was on the mathematics test in which the student scored .75 standard deviation units above the mean. Note that although the student had the same score of 42 on the science and social studies tests, the score was above the mean in social studies but below the mean in science. The following frequency curve illustrates these comparisons:



Inferential Statistics

Researchers use inferential statistics to make inferences about a population of study participants based on a sample of these participants. For example, a researcher might attempt to conclude something about a population of students (e.g., all 4th graders in a school district) by studying a sample of these students. Based on inferential statistics, the researcher infers that the results from the sample of 4th graders are also true of the population of 4th graders. Inferential statistics also are used to make inferences about the differences between two or more groups of observations.

Example 3:
A researcher randomly selects participants from a population of 4th-grade students and randomly assigns them to two groups. Students in Group A participate in Reading Program A. Students in Group B participate in Reading Program B. Based on their reading test scores, which program resulted in better reading performance?

Program A:
Scores = 14,11,15,15,16,16,18,18,17,18,14,17,12,15,16,14,15,13
n = 18
M = 15.22
SD = 2.02

Program B:
Scores = 15,22,19,20,22,20,21,14,20,21,19,19,16,12,18,17,20,18
n = 18
M = 18.5
SD = 2.77

Reading Scores	Frequency	Percent	Percentile
PROGRAM A			
11	1	5.6	5.6
12	1	5.6	11.1
13	1	5.6	16.7
14	3	16.7	33.3
15	4	22.2	55.6
16	3	16.7	72.2
17	2	11.1	83.3
18	3	16.7	100.0
PROGRAM B			
12	1	5.6	5.6
14	1	5.6	11.1
15	1	5.6	16.7
16	1	5.6	22.2
17	1	5.6	27.8
18	2	11.1	38.9
19	3	16.7	55.6
20	4	22.2	77.8
21	2	11.1	89.9
22	2	11.1	100.0

According to the descriptive statistics and the frequency graph, Program B resulted in better reading performance because students in Group B achieved a higher mean test score than students in Group A. Is this difference however of 3.28 between the means of the two groups due to Program B, or could this difference simply be due to chance factors? To answer this question requires the use of inferential statistics.

STATISTICAL SIGNIFICANCE

The research design of the study determines the type of inferential statistic used. All inferential statistics however answer the same question:

Could these findings occur by chance or are these findings too unlikely to occur by chance and therefore the findings reflect a real effect of what is being studied?

The most common inferential statistics are the *t*-test and the *F*-test (also known as analysis of variance). The *t* statistic is used when there are two groups of participants in the research study. The *F* statistic is used when there are more than two groups in the research study. Usually, the researcher uses a computer program to calculate the inferential test statistic and the probability of obtaining a particular statistical value if there is no real difference between the groups.

In Example 3, the *t* statistic is 4.06. The researcher would report this result as follows:
Students in Group B performed significantly better than students in Group A, $t = 4.06$ (34) $p < .001$. What does this mean?

Simply put, the probability of this result occurring by chance is less than one time out of 1,000. Therefore, the researcher can be very confident that the difference between the two groups reflects an actual difference. [Note: The number 34 in parentheses is called the degrees of freedom and reflects the size of the samples. For a two-sample *t*-test, the degrees of freedom are calculated as $(n - 1) + (n - 1)$. Degrees of freedom are used in the calculation of inferential statistics, and it is conventional to report them.]

The term statistically significant is used to describe results for which there is a 5% or less probability that the results occurred by chance. Why 5%? By convention, social scientists have chosen this percentage as the cut-off point (although other percentages are sometimes chosen). Therefore, any result that has a probability of occurring by chance more than five times out of 100 (designated by convention as $p > .05$) is reported as not significant. Researchers should not discuss nonsignificant results as if they indicate actual differences between groups.

Sometimes researchers also report the confidence interval for the results of a *t*-test. In Example 2, the 95% confidence interval for the mean difference between Programs A and B is between 1.63 and 4.92. This means that if the entire population of 4th-grade students participated in the two reading programs, there is a 95% probability that the mean difference in reading achievement between Programs A and B would be between 1.63 and 4.92 points. The confidence interval provides an estimate of population measurements based on sample measurements.

There is an important relationship between the size of the sample and statistical significance. As the sample size increases, the probability increases that significant differences will be detected. This is a concept called statistical power.

Consider results from the following studies:

Example 4:
Program X: n = 10, Mean achievement = 30.5
Program Y: n = 10, Mean achievement = 31.5
$t = 2.15$, $p > .05$
The difference between Program X and Program Y *is not statistically significant*.

Example 5:
Program X: n = 100, Mean achievement = 30.5
Program Y: n = 100, Mean achievement = 31.5
$t = 2.15$, $p < .05$

The difference between Program A and Program B is *statistically significant*.

The same numerical difference of 1.5 points between the two groups is statistically significant in the study with large sample sizes (and more statistical power) but not in the study with small sample sizes. In studies with very large sample sizes (e.g., 1,000), even small numerical differences can be statistically significant. For this reason, it is important to examine what is known as the effect size of a statistically significant difference.

PRACTICAL SIGNIFICANCE

In addition to measures of statistical significance, researchers frequently calculate and report measures of practical significance, known as the effect size. The effect size helps policymakers and educators decide whether a statistically significant difference between programs translates into enough of a difference to justify adoption of a program.

There are different ways to measure effect sizes. One commonly used measure is called Cohen's *d*, which measures effect sizes in standard deviation units. In Example 3, Cohen's $d = 1.34$ standard deviation units. Social scientists commonly interpret *d* as follows (although interpretation also depends on the intervention and the dependent variable):

- Small effect sizes: $d = .2$ to $.5$
- Medium effect sizes: $d = .5$ to $.8$
- Large effect sizes: $d = .8$ and higher

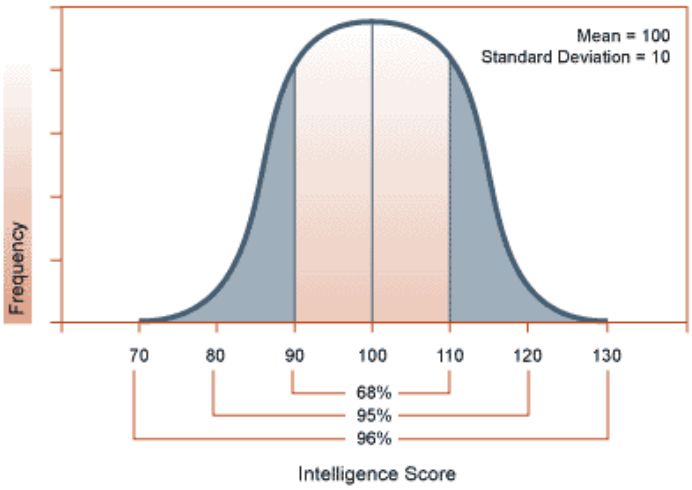
Thus, in Example 3, the effect size of $d = 1.34$ is "large," but what does "large" mean in terms of reading achievement?

A simple way to understand effect sizes is to translate d into percentile gains. An effect size of $d = 1.34$ translates into a percentile gain of 41 percentile points (based on the normal curve, as described in the next section). This means that the reading score of the average student who participates in Reading Program B will be 41 percentile points higher than the average student who participates in Reading Program A. The bottom line: Program B is a more effective reading program than Program A.

THE NORMAL CURVE AND EFFECT SIZES

Another way to understand effect sizes is to examine the normal curve. The normal curve refers to a frequency distribution in which the graph of scores resembles a bell — hence, the famous bell-shaped curve. Many human traits such as intelligence, personality scores and student achievement have normal distributions.

Example 6:
If all adults in the state of Colorado were given a general intelligence test, the frequency distribution of the scores would resemble the following bell-shaped curve.



The normal distribution has an important characteristic. The mean, median and mode are the same score (a score of 100 in Example 6) because a normal distribution, is symmetrical. The score with the highest frequency occurs in the middle of the distribution and exactly half of the scores occur above the middle and half of the scores occur below. Most of the scores occur around the middle of the distribution or the mean. Very high and very low scores occur infrequently and are therefore considered rare.

In a normal distribution, 34.1 % of the scores occur between the mean and one standard deviation above the mean. In Example 6, the standard deviation is 10. The result is that 34.1% of adults in Colorado scored between 100 and 110. (Conversely, 34.1% of adults in Colorado scored between 100 and 90.) A score of 120 is two standard deviations above the mean. In a normal distribution, 47.5% of the scores occur between the mean and two standard deviations above or below the mean. Thus, two standard deviations above and below the mean include 95% of all scores.

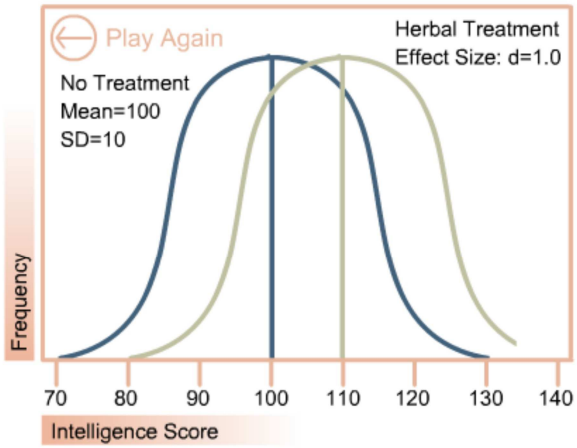
Scores in a normal distribution also can be described as percentiles. The score that is the mean (and also the median and mode) is the score at the 50th percentile because 50% of the scores are at that score or below. In the example, a score of 100 is at the 50th percentile. A score of 110 is one standard deviation above the mean and therefore at the 84th percentile (50% + 34.1%). Finally, a score of 120 is two standard deviations above the mean and is therefore at the 97th percentile (50% + 47.5%).

Hint:
Sometimes percentile scores on tests are converted into normal curve equivalent (NCE) scores because NCE scores are easier to manipulate arithmetically and statistically than are percentiles.

How do effect sizes relate to the normal curve? Because Cohen's *d* is measured in standard deviation units, an effect size of *d* = 1.0 is equal to one standard deviation above the mean.

Example 7:
A researcher discovers a special herb that increases adult intelligence, with an effect size of *d* = 1.0. The average adult in Colorado (with an intelligence score of 100) who takes this herb can expect to have an intelligence score of 110, an increase in

percentile rank from the 50th percentile to the 84th percentile. This researcher stands to make a lot of money!



Effect sizes also apply to scores on student achievement tests because these tests are designed to be normally distributed. For example, an effect size of d = 1.0 for a reading program means that the reading program increased the reading score of the average student to one standard deviation above the mean. An effect size of d = .5 means that the reading score of the average student in the program increased to .5 standard deviation above the mean. (If the standard deviation equals 8, the average student's score would increase by 8 points with d = 1.0, and would increase by 4 points with d = .5.)

Caution:
Effect sizes also can be negative, which means that scores are lowered by the effect of the program in the study. For example, an effect size of d = –1.0 means that the average score was decreased by one standard deviation.

Correlation

Correlation refers to a technique used to measure the relationship between two or more variables.

Example 8:
In the following example, the first variable is the number of students in 4th-grade classes in a school district. The second variable is the mean reading score of each class.

VARIABLE 1 Class Size	VARIABLE 2 Mean Reading Score
25	70
20	80
25	60
25	72
30	58
22	71
28	68
20	75
19	72
29	61

Pearson *r* is a statistic that is commonly used to calculate bivariate correlations. In Example 8, Pearson $r = -.80$, $p < .01$. What does this mean?

To interpret correlations, four pieces of information are necessary.

1. *The numerical value of the correlation coefficient.* Correlation coefficients can vary numerically between 0.0 and 1.0. The closer the correlation is to 1.0, the stronger the relationship between the two variables. A correlation of 0.0 indicates the absence of a relationship. In Example 8, the correlation coefficient is –.80, which indicates the presence of a strong relationship.

2. *The sign of the correlation coefficient.* A positive correlation coefficient means that as variable 1 increases, variable 2 increases, and conversely, as variable 1 decreases, variable 2 decreases. In other words, the variables move in the same direction when there is a positive correlation. A negative correlation means that as variable 1 increases, variable 2 decreases and vice versa. In other words, the variables move in opposite directions when there is a negative correlation. In Example 8, the negative sign indicates that as class size increases, mean reading scores decrease.

3. *The statistical significance of the correlation.* A statistically significant correlation is indicated by a probability value of less than .05. This means that the probability of obtaining such a correlation coefficient by chance is less than five times out of 100, so the result indicates the presence of a relationship. In Example 8, there is a statistically significant negative relationship between class size and reading score ($p < .001$), such that the probability of this correlation occurring by chance is less than one time out of 1000.

4. *The effect size of the correlation.* For correlations, the effect size is called the coefficient of determination and is defined as r^2. The coefficient of determination can vary from 0 to 1.00 and indicates that the proportion of variation in the scores can be predicted from the relationship between the two variables. In Example 8, the coefficient of determination is .65, which means that 65% of the variation in mean reading scores among the different classes can be predicted from the

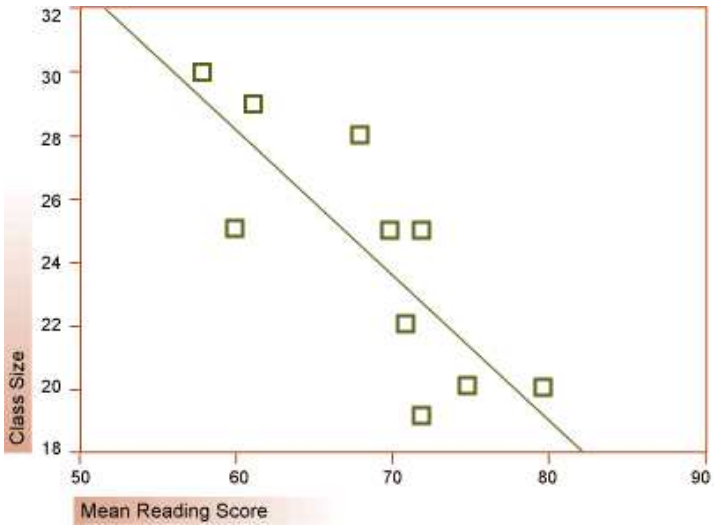
relationship between class size and reading scores. (Conversely, 35% of the variation in mean reading scores cannot be explained.)

A correlation can only indicate the presence or absence of a relationship, not the nature of the relationship. In Example 8, it cannot be concluded that smaller class sizes cause higher reading scores, even if the correlation is 1.0. *Correlation is not causation*. There is always the possibility that a third variable influenced the results. For example, perhaps the students in the small classes were higher in verbal ability than the students in the large classes or were from higher income families or had higher quality teachers.

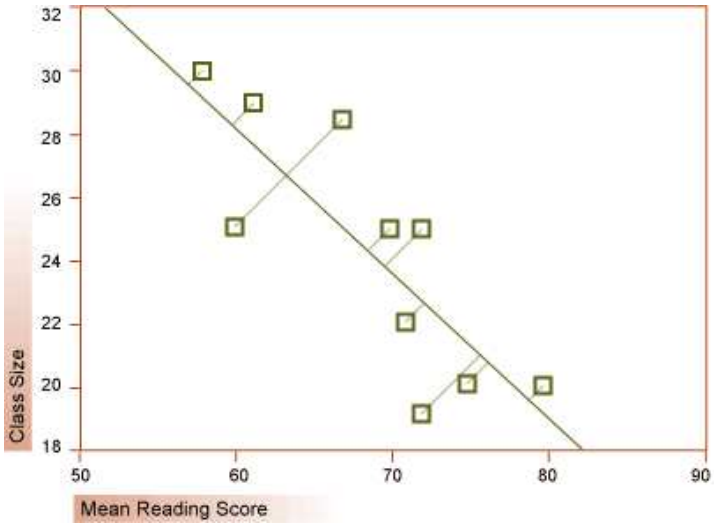
CORRELATION AND PREDICTION

Another use of correlation is prediction. A mathematical technique called regression analysis uses the correlation between two variables to predict the values for variable 2 (the dependent or criterion variable) based on the values for variable 1 (the predictor variable).

The following graph indicates a linear relationship between variable 1 and variable 2 from Example 8.



A regression analysis can identify the equation that best describes the linear relationship between class size and reading score in the graph. This equation can then be used to estimate mean reading scores based on class sizes. Unless there is a perfect correlation between two variables (i.e., $r = \pm 1.00$), the prediction based on regression analysis will be imperfect. The standard error of estimate indicates how accurately the equation can predict values of a variable. In the example, the standard error of estimate is 4.44, which is the average distance between the line in a graph of the regression equation and the actual data points for the mean reading score.



A simple way to think about prediction error is that the smaller the numerical value of the correlation, the smaller the coefficient of determination, and the more error there will be when using the correlation for prediction.

CORRELATION WITH MULTIPLE VARIABLES

When there is more than one predictor variable, the technique of multiple regression analysis combines the predictor variables to produce a multiple correlation coefficient called *R*. For example, in addition to class size, a researcher might use students' mean verbal ability scores and socioeconomic status to predict reading scores. A multiple correlation coefficient of $R = .71$ would indicate the degree of the combined correlation of the predictor variables with mean reading scores. The squared multiple correlation coefficient of $R^2 = .49$ would indicate that 49% of the variation among mean reading scores of the different 4th-grade classes can be predicted by the relationship between reading scores and the combination of class size, verbal ability and socioeconomic status. (Conversely, 51% of the variation in mean reading scores cannot be explained.)

Although the technique of multiple regression provides more information than bivariate correlation, it cannot be concluded that variables caused other variables to occur in certain ways.

STRUCTURAL EQUATION MODELING

Like multiple regression, structural equation modeling (SEM) also examines linear relationships among a set of variables. With SEM however, the researcher hypothesizes a model for how the variables that are measured in a study are related to one another as well as how the measured variables influence and are influenced by unobserved variables called latent variables. For example, student motivation might be a latent variable that influences student achievement and class size might influence student motivation. In SEM, the statistics that are of primary interest are goodness-of-fit statistics that evaluate how well the data fit the researcher's proposed model for the interrelationships among the variables.

Show me a Structural Equation Model

Caution:
Structural equation modeling is sometimes referred to as causal path modeling. Despite the use of the word "causal," this technique is correlational and does not support conclusions about cause and effect.

HIERARCHICAL LINEAR MODELING

Hierarchical Linear Modeling (HLM) is statistical technique used when the data are from participants who exist within different levels of a hierarchical structure (Osborne, 2000). For example, students exist within a hierarchical structure that includes family, classroom, grade, school, district and state. Student achievement is considered nested data because it reflects influences from each of these levels (e.g., influences from family characteristics; the classroom teacher; the grade level; and school, district and state policies).

With HLM, the researcher first measures the influence of one or more predictor variables (e.g., student socioeconomic status and prior achievement) on an outcome (student reading achievement) at level one. Next the researcher measures the relationship of level two predictor variables (e.g., teacher professional development and experience) on the level one relationship. For example, through HLM, a researcher might find that student socioeconomic status and prior achievement are negatively related to reading achievement, but that this relationship is less strong with increasing teacher professional development. In other words, the more professional development teachers have, the weaker the correlation of these other factors is with their students' achievement.

Reference and Resources

Osborne, J. W. (2000). Advantages of hierarchical liner modeling. *ERIC/AE Digest*. (ERIC Document Reproduction Service No. ED447198)

Gravetter, F. J. & Wallnau, L. B. (1988). *Statistics for the behavioral sciences* (2nd ed.). St. Paul: West Publishing.

Grimm, L. G. & Yarnold, P. R. (Eds.). (1995). *Reading and understanding multivariate statistics*. Washington DC: American Psychological Association.

Maruyama, G. M. (1998). *Basics of structural equation modeling*. Thousand Oaks: Sage Publications