Tutorials in Quantitative Methods for Psychology
2007, vol. 3 (2), p. 43-50.

DOI: 10.20982/tqmp.03.2.p043

Understanding Power and Rules of Thumb for Determining Sample Sizes

Carmen R. Wilson VanVoorhis and Betsy L. Morgan
University of Wisconsin-La Crosse

This article addresses the definition of power and its relationship to Type I and Type II errors. We discuss the relationship of sample size and power. Finally, we offer statistical rules of thumb guiding the selection of sample sizes large enough for sufficient power to detecting differences, associations, chi-square, and factor analyses.

As researchers, it is disheartening to pour time and intellectual energy into a research project, analyze the data, and find that the elusive .05 significance level was not met. If the null hypothesis is genuinely true, then the findings are robust. But, what if the null hypothesis is false and the results failed to detect the difference at a high enough level? It is a missed opportunity. Power refers to the probability of rejecting a false null hypothesis. Attending to power during the design phase protect both researchers and respondents. In recent years, some Institutional Review Boards for the protection of human respondents have rejected or altered protocols due to design concerns (Resnick, 2006). They argue that an "underpowered" study may not yield useful results and consequently unnecessarily put respondents at risk. Overall, researchers can and should attend to power. This article defines power in accessible ways, provides guidelines for increasing power, and finally offers "rules-of-thumb" for numbers of respondents needed for common statistical procedures.

What is power?

Beginning social science researchers learn about Type I and Type II errors. Type I errors (represented by α are made when the data result in a rejection of the null hypothesis, but in reality the null hypothesis is true (Neyman & Pearson (1928/1967). Type II errors (represented by β) are made when the data do not support a rejection of the null hypothesis, but in reality the null hypothesis is false (Neyman & Pearson). However, as shown in Figure 1, in every study, there are four possible outcomes. In addition to Type I and Type II errors, two other outcomes are possible. First, the data may not support a rejection of the null hypothesis when, in reality, the null hypothesis is true. Second, the data may result in a rejection of the null hypothesis when, in reality, the null hypothesis is false (see Figure 1). This final outcome represents statistical power. Researchers tend to over-attend to Type I errors (e.g., Wolins, 1982), in part, due to the statistical packages that rarely include estimates of the other probabilities. Post-hoc analyses of published articles often yield the finding that Type II errors are common events in published articles (e.g., Strasaik, Zamanm, Pfeiffer, Goebel, & Ulmer, 2007; Williams, Hathaway, Kloster, & Layne, 1997).

When a .05 or lower significance is obtained, researchers are fairly confident that the results are "real," in other words not due to chance factors alone. In fact, with a significance level of .05, researchers can be 95% confident the results represent a non-chance finding (Aron & Aron, 1999). Researchers should continue to strive to reduce the probability of Type I errors; however, they also need to increase their attention to power.

Every statistic has a corresponding sampling distribution. A sampling distribution is created, in theory, via the following steps (Kerlinger & Lee, 2000):

1. Select a sample of a given *n* under the null hypothesis.
2. Calculate the specific statistic.
3. Repeat steps 1 and 2 an "infinite" number of times.

Portions of this article were published in *Psi Chi Journal of Undergraduate Research*.

Figure 1. Possible outcomes of decisions based on statistical results.

		"TRUTH" OR "REALITY"	
		Null correct	*Null wrong*
Decision based on statistical result	*Fail to reject*	Correct decision	Type II (β)
	Reject	Type I (α)	Correct decision **Power**

4. Plot the given statistic by frequency of value.

For instance, the following steps could be used to create a sampling distribution for the independent samples *t*-test (based on Fisher 1925/1990; Pearson, 1990).

1. Select two samples of a given size from a single population. The two samples are selected from a single population because the sampling distribution is constructed given the null hypothesis is true (i.e., the sample means are not statistically different).
2. Calculate the independent samples *t*-test statistic based on the two samples.
3. Complete steps 1 and 2 an "infinite" number of times. In other words, select two samples from the same population and calculate the independent samples *t*-test statistic repeatedly.
4. Plot the obtained independent samples *t*-test values by frequency. Given the independent samples *t*-test is based on the difference between the means of the two samples, most of the values will hover around zero as the samples both were drawn from the same population (i.e. both sample means are estimating the same population mean). Sometimes, however, one or both of the sample means will be poor estimates of the population mean and differ widely from each other, yielding the bell-shaped curve characteristic of the independent samples *t*-test sampling distribution.

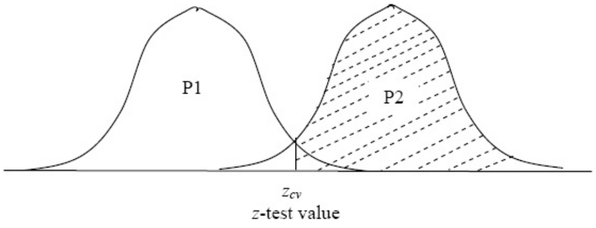

Figure 2. Sampling distributions of means for the *z*-test assuming the null hypothesis is false. P1 represents the sampling distribution of means of the original population; P2 represents the sampling distribution of means from which the sample was drawn. The shaded area under P2 represents power, i.e., the probability of correctly rejecting a false null hypothesis.

When a researcher analyzes data and calculates a statistic, the obtained value is compared against this sampling distribution. Depending on the location of the obtained value along the sampling distribution, one can determine the probability of achieving that particular value given the null hypothesis is true. If the probability is sufficiently small, the researcher rejects the null hypothesis. Of course, the possibility remains, albeit unlikely, that the null hypothesis is true and the researcher has made a Type I error.

Estimating power depends upon a different distribution (Cohen, 1992). The simplest example is the z-test, in which the mean of a sample is compared to the mean of the population to determine if the sample comes from the population (P1). Power assumes that the sample, in fact, comes from a different population (P2). Therefore, the sampling distribution of P2 will be different than the sampling distribution of P1 (see Figure 2). Power assumes that the null hypothesis is *incorrect*.

The goal is to obtain a z-test value sufficiently extreme to reject the null hypothesis. Usually, however, the two distributions overlap. The greater the overlap, the more values P1 and P2 share, and the less likely it is that the obtained test value will result in the rejection of the null hypothesis. Reducing this overlap increases the power. As the overlap decreases, the proportion of values under P2 which fall within the rejection range (indicated by the shaded area under P2) increases.

Table 1 : Sample Data Set

Person	X	Person	X
1	5.50	11	7.50
2	6.00	12	7.50
3	6.00	13	8.00
4	6.50	14	8.00
5	6.50	15	8.00
6	7.00	16	8.50
7	7.00	17	8.50
8	7.00	18	9.00
9	7.50	19	9.00
10	7.50	20	9.50

Manipulating Power

Sample Sizes and Effect Sizes

As argued earlier a reduction of the overlap of the distributions of two samples increases power. Two strategies exist for minimizing the overlap between distributions. The first, and the one a researcher can most easily control, is to increase the sample size (e.g., Cohen, 1990; Cohen, 1992). Larger samples result in increased power. The second, discussed later, is to increase the effect size.

Larger samples more accurately represent the characteristics of the populations from which they are derived (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Marcoulides, 1993). In an oversimplified example, imagine a population of 20 people with the scores on some measure (X) as listed in Table 1.

The mean of this "population" is 7.5 (σ = 1.08). Imagine researchers are unable to know the exact mean of the population and wanted to estimate it via a sample mean. If they drew a random sample, n = 3, it could be possible to select three low or three high scores which would be rather poor estimates of the "population" mean. Alternatively, if they drew samples, n = 10, even the ten lowest or ten highest scores would better estimate the population mean than the sample of three. For example, using this "population" we drew ten samples of three and ten samples of ten (see Table 2 for the sample means).

The overall mean of the sample means based on three people is 7.57 and the standard deviation is .45. The overall mean of the sample means based on ten people is 7.49 and the standard deviation is .20. The sample means based on ten people were, on average, closer to the population mean (μ = 7.50) than the sample means based on three people.

The standard error of measurement estimates the average difference between a sample statistic and the population statistic. In general, the standard error of measurement is the standard deviation of the sampling distribution. In the above example, we created two miniature sampling distributions of means. The sampling distribution of the z-test (used to compare a sample mean to a population mean) is a sampling distribution of means (although it includes an "infinite" number of sample means). As indicated by the standard deviations of the means (i.e., the standard error of measurements) the average difference between the sample means and the population mean is smaller when we drew samples of 10 than when we drew samples of 3. In other words, the sampling

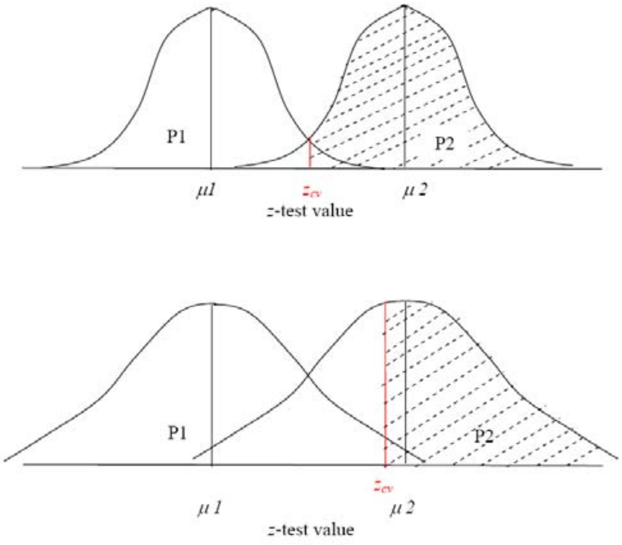

Figure 3. The relationship between standard error of measurement and power. As the standard error of measurement decreases, the proportion of the P2 distribution above the z-critical value (see shaded area under P2) increases, therefore increasing the power. The distributions at the top of the figure have smaller standard errors of measurement and therefore less overlap, while the distributions at the bottom have larger standard errors of measurement and therefore more overlap, decreasing the power.

distribution based on samples of size 10 is "narrower" than the sampling distribution based on samples of size 3. Applied to power, given the population means remain static, "narrower" distributions will overlap less than "wider" distributions (see Figure 3).

Consequently, larger sample sizes increase power and decrease estimation error. However, the practical realities of conducting research such as time, access to samples, and financial costs restrict the size of samples for most researchers. The balance is generating a sample large enough to provide sufficient power while allowing for the ability to actually garner the sample. Later in this article, we provide some "rules of thumb" for some common statistical tests aimed at obtaining this balance between resources and ideal sample sizes.

The second way to minimize the overlap between distributions is to increase the effect size (Cohen, 1988). Effect size represents the actual difference between the two populations; often effect sizes are reported in some standard unit (Howell, 1997). Again, the simplest example is the *z*-test. Assuming the null hypothesis is false (as power does), the effect size (*d*) is the difference between the $\mu 1$ and $\mu 2$ in standard deviation units. Specifically,

$$d = \frac{M - \mu_1}{\sigma_1}$$

where *M* is the sample mean derived from $\mu 2$ (remember, power assumes the null hypothesis is false, therefore, the sample is drawn from a different population than $\mu 1$.) If the effect size is .50, then $\mu 1$ and $\mu 2$ differ by one-half of a standard deviation. The more disparate the population means, the less overlap between the distributions (see Figure 4). Researchers can increase power by increasing the effect size.

Manipulating effect size is not nearly as straightforward as increasing the sample size. At times, researchers can attempt to maximize effect size by maximizing the difference between or among independent variable levels. For example, suppose a particular study involved examining the effect of caffeine on performance. Likely differences in performance, if they exist, will be more apparent if the researcher compares individuals who ingest widely different amounts of caffeine (e.g., 450 mg vs. 0 mg) than if she compares individuals who ingest more similar amounts of caffeine (e.g., 25 mg. vs. 0 mg). If the independent variable is a measured subject variable, for example, ability level, effect size can be increased by including groups who are "extreme" in ability level. For example, rather than

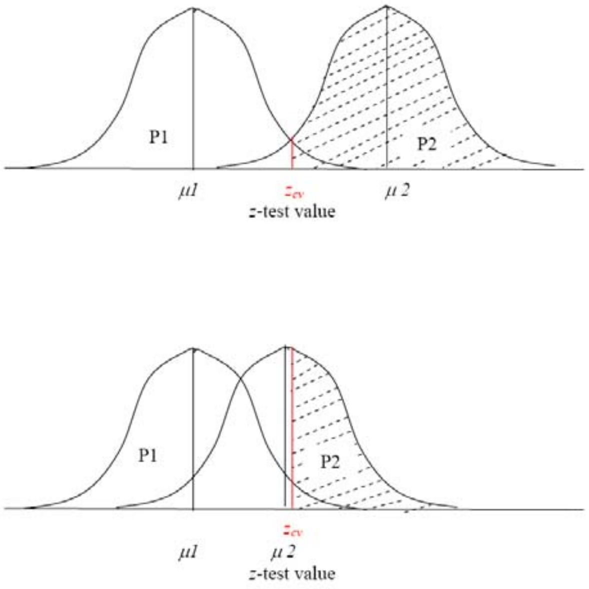

Figure 4. The relationship between effect size and power. As the effect size increases, the proportion of the P2 distribution above the z-critical value (see shaded area under P2) increases, therefore increasing the power. The distributions at the top of the figure represent populations with means that differ to a larger degree (i.e. a larger effect size) than the distributions at the bottom. The larger difference between the population means results in less overlap between the distributions, increasing power.

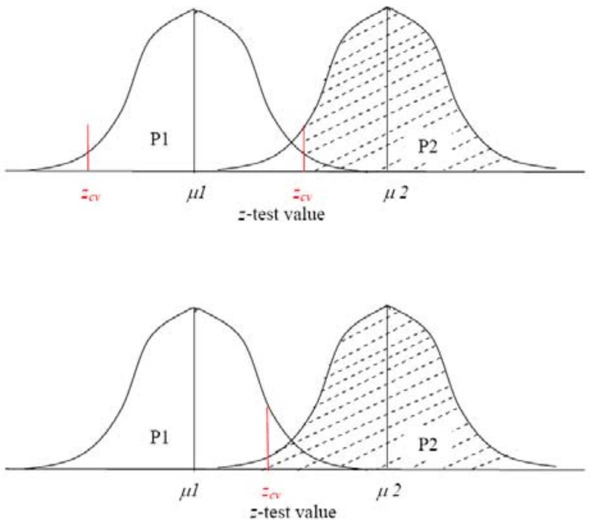

Figure 5. The relationship between α and power. As α increases, as in a single-tailed test, the proportion of the P2 distribution above the *z* critical value (see shaded area under P2). The distributions at the top of the figure represent a two-tailed test in which the α level is split between the two tails; the distributions at the bottom of the figure represent a one-tailed test in which the α level is included in only one tail.

Table 2. *Sample Means Presented by Magnitude*

Sample	M (n = 3)	M (n = 10)
1	6.83	7.25
2	7.17	7.30
3	7.33	7.30
4	7.33	7.35
5	7.50	7.40
6	7.67	7.50
7	7.67	7.60
8	7.83	7.70
9	7.83	7.75
10	8.50	7.75

comparing people who are above the mean in ability level with those who are below the mean, the researcher might compare people who score at least one standard deviation above the mean with those who score at least one standard deviation below the mean. Other times, the effect size is simply out of the researcher's control. In those instances, the best a researcher can do is to be sure the dependent variable measure is as reliable as possible to minimize any error due to the measurement (which would serve to "widen" the distribution).

Error Variance and Power

Error variance, or variance due to factors other than the independent variable, decreases the likelihood of detecting differences or relationships that actually exist, i.e. decreases power (Cohen, 1988). Differences in dependent variable scores can be due to many factors other than the effects of the independent variable. For example, scores on measures with low reliability can vary dependent upon the items included in the measure, the conditions of testing, or the time of testing. A participant might be talented in the task or, alternatively, be tired and unmotivated. Dependent samples control for error variance due to such participant characteristics.

Each participant's dependent variable score (X) can be characterized as

$$X = \mu + t_x + p + e$$

where μ is the population mean, t_x is the effects of treatment, p is the participant characteristics, and e is random error.

In the true dependent samples design, each participant experiences each level of the independent variable. Any participant characteristics which impact the dependent variable score at one level will similarly affect the dependent variable score at other levels of the independent variable. Different statistics use different methods to separate variance due to participant characteristics from error variance. The simplest example is a dependent samples *t*-test design, in which there are two levels of an independent variable. The formula for the dependent samples *t*-test is

$$t = \frac{MD}{SE_{MD}}$$

where *MD* is the mean of the difference scores and SE_{MD} is the standard error of the mean difference.

Difference scores are created for each participant by subtracting the score under one level of the independent variable from the score under the other level of the independent variable. The actual magnitude of the scores, then is eliminated, leaving a difference that is due, to a larger degree, to the treatment and to a lesser degree to participant characteristics. The differences due to treatment then are easier to detect. In other words, such a design increases power (Cohen, 2001; Cohen, 1988).

Type I Errors and Power

Finally, power is related to α,or the probability of making a Type I error. As α increases, power increases (see Figure 5). The reality is that few researchers or reviewers are willing to trust in results where the probability of rejecting a true null hypothesis is greater than .05. Nonetheless, this relationship does explain why one-tailed tests are more powerful than two-tailed tests. Assuming an α level of .05, in a two-tailed test, the total α level must be split between the tails, i.e., .025 is assigned to each tail. In a one-tailed test, the entire α level is assigned to one of the tails. It is as if the α level has increased from .025 to .05.

Rules of Thumb

The remaining articles in this edition discuss specific power estimates for various statistics. While we certainly advocate for full understanding of and attention to power estimates, at times, such concepts are beyond the scope of a particular researchers training (for example, in undergraduate research). In those instances, power need not be ignored totally, but rather can be attended to via certain rules of thumb based on the principles of regarding power. Table 3 provides an overview of the sample size rules of thumb discussed below.

Table 3: *Sample size rules of thumb*

Relationship	Reasonable sample size
Measuring group differences (e.g., *t*-test, ANOVA)	Cell size of 30 for 80% power, if decreased, no lower than 7 per cell.
Relationships (e.g., correlations, regression)	~50
Chi - Square	At least 20 overall, no cell smaller than 5.
Factor Analysis	~300 is “good”

Number of Participants: Cell size for statistics used to detect differences.

The independent samples *t*-test, matched sample *t*-test, ANOVA (one-way or factorial), MANOVA are all statistics designed to detect differences between or among groups. How many participants are needed to maintain adequate power when using statistics designed to detect differences? Given a medium to large effect size**,** *30 participants per cell* should lead to about 80% power (the minimum suggested power for an ordinary study) (Cohen, 1988). Cohen conventions suggest an effect size of .20 is small, .50 is medium, and .80 is large. If, for some reason, minimizing the number of participants is critical, 7 participants per cell, given at least three cells, will yield power of approximately 50% when the effect size is .50. Fourteen participants per cell, given at least three cells and an effect size of .50, will yield power of approximately 80% (Kraemer & Thiemann, 1987).

Caveats. First, comparisons of fewer groups (i.e., cells) require more participants to maintain adequate power. Second, lower expected effect sizes require more participants to maintain adequate power (Aron & Aron, 1999). Third, when using MANOVA, it is important to have more cases than dependent variables (DVs) in every cell (Tabachnick & Fidell, 1996).

Number of participants: Statistics used to examine relationships.

Although there are more complex formulae, the general rule of thumb is no less than 50 participants for a correlation or regression with the number increasing with larger numbers of independent variables (IVs). Green (1991) provides a comprehensive overview of the procedures used to determine regression sample sizes. He suggests $N > 50 + 8$ m (where m is the number of IVs) for testing the multiple correlation and $N > 104 + \text{m}$ for testing individual predictors (assuming a medium-sized relationship). If testing both, use the larger sample size.

Although Green's (1991) formula is more comprehensive, there are two other rules of thumb that could be used. With five or fewer predictors (this number would include correlations), a researcher can use Harris's (1985) formula for yielding the absolute minimum number of participants. Harris suggests that the number of participants should exceed the number of predictors by at least 50 (i.e., total number of participants equals the number of predictor variables plus 50)--a formula much the same as Green's mentioned above. For regression equations using six or more predictors, an absolute minimum of 10 participants per predictor variable is appropriate. However, if the circumstances allow, a researcher would have better power to detect a small effect size with approximately 30 participants per variable. For instance, Cohen and Cohen (1975) demonstrate that with a single predictor that in the population correlates with the DV at .30, 124 participants are needed to maintain 80% power. With five predictors and a population correlation of .30, 187 participants would be needed to achieve 80% power.

Caveats. Larger samples are needed when the DV is skewed, the effect size expected is small, there is substantial measurement error, or stepwise regression is being used (Tabachnick & Fidell, 1996).

Number of participants: Chi-square.

The chi-square statistic is used to test the independence of categorical variables. While this is obvious, sometimes the implications are not. The primary implication is that all observations must be independent. In other words, no one individual can contribute more than one observation. The degrees of freedom are based on the number of variables and their possible levels, not on the number of observations. Increasing the number of observations, then has no impact on the critical value needed to reject the null hypothesis.

The number of observations still impacts the power, however. Specifically, *small expected frequencies* in one or more cells limit power considerably. Small expected frequencies can also slightly inflate the Type I error rate, however, for totally sample sizes of at least 20, the alpha rarely rises above .06 (Howell, 1997). A conservative rule is that no expected frequency should drop below 5.

Caveat. If the expected effect size is large, lower power can be tolerated and total sample sizes can include as few as 8 observations without inflating the alpha rate.

Number of Participants: Factor analysis.

A good general rule of thumb for factor analysis is 300 cases (Tabachnick & Fidell, 1996) or the more lenient 50 participants per factor (Pedhazur & Schmelkin, 1991). Comrey and Lee (1992) (see Tabachnick & Fidell, 1996) give the following guide samples sizes: 50 as very poor; 100 as poor, 200 as fair, 300 as good, 500 as very good and 1000 as excellent.

Caveat. Guadagnoli & Velicer (1988) have shown that solutions with several high loading marker variables (>.80) do not require as many cases.

Conclusion

This article addresses the definition of power and its relationship to Type I and Type II errors. Researchers can manipulate power with sample size. Not only does proper sample selection improve the probability of detecting difference or association, researchers are increasingly called upon to provide information on sample size in their human respondent protocols and manuscripts (including effect sizes and power calculations). The provision of this level of analysis regarding sample size is a strong recommendation of the Task Force on Statistical Inference (Wilkinson, 1999), and is now more fully elaborated in the discussion of "what to include in the Results section" of the new fifth edition of the American Psychological Association's (APA) publication manual (APA, 2001). Finally, researchers who do not have the access to large samples should be alert to the resources available for minimizing this problem (e.g., Hoyle, 1999).

References

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Aron, A., & Aron, E. N. (1999). *Statistics for psychology* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Cohen, B. H. (2001). *Explaining Psychological Statistics* (2nd ed.). New York, NY: John Wiley & Sons, Inc.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304-1312.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155-159.

Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cronbach, L. J. , Gleser, G. C., Nanda, H., & Rajaratnam, N., (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Fisher, R. A. (1925/1990). *Statistical methods for research workers*. Oxford, England: Oxford University Press.

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research, 26,* 499-510.

Guadagnoli, E., & Velicer, W.F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin, 103,* 265-275.

Harris, R. J. (1985). *A primer of multivariate statistics* (2nd ed.). New York: Academic Press.

Howell, D. C. (1997). *Statistical methods for psychology* (4th ed.). Belmont, CA: Wadsworth.

Hoyle, R. H. (Ed.). (1999). *Statistical strategies for small sample research.* Thousand Oaks, CA: Sage.

Kerlinger, F. & Lee, H. (2000). *Foundations of behavioral research.* New York: International Thomson Publishing.

Kraemer, H. C., & Thiemann, S. (1987). *How many subjects? Statistical power analysis in research.* Newbury Park, CA: Sage.

Marcoulides, G. A. (1993). Maximizing power in generalizability studies under budget constraints. Journal of Educational Statistics, 18 (2), 197-206.

Neyman, J. & Pearson, E. S. (1928/1967). On the use and interpretation of certain test criteria for purposes of statistical inference, Part I. *Joint Statistical Papers*. London: Cambridge University Press.

Pearson , E, S. (1990) *'Student', A statistical biography of William Sealy Gosset.* Oxford, England: Oxford University Press.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Erlbaum.

Resnick, D. B. (2006, Spring) *Bioethics bulletin.* Retrieved September 22, 2006 from http://dir.niehs.nih.gov/ethics/news/2006spring.doc. Washington DC: National Institute for Environmental Ethics Health Sciences.

Strasaik, A. M, Zamanm, Q., Pfeiffer, K. P., Goebel, G., Ulmer, H. (2007). Statistical errors in medical research: A review of common pitfalls. Swiss Medical Weekly, 137, 44-49.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York: HarperCollins.

Wilkinson, L., & Task Force on Statistical Inference, APA Board of Scientific Affairs. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604.

Williams, J. L., Hathaway, C. A., Kloster, K. L. & B. H. Layne, (1997). Low power, type II errors, and other statistical problems in recent cardiovascular research. *Heart and Circulatory Physiology, 273,* (1). 487-493.

Wolins, L. (1982). Research mistakes in the social and behavioral sciences. Ames: Iowa State University Press

Manuscript received October 21st, 2006
Manuscript accepted November 5th, 2007