Practical Assessment, Research, and Evaluation

Volume 9 *Volume 9, 2004* Article 3

2004

The Contributions of Reliability and Pretests to Effective Assessment

Donald Bacon

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 9, Number 3, February, 2004 ISSN=1531-7714

The Contributions of Reliability and Pretests to Effective Assessment

Donald R. Bacon
University of Denver

This paper shows how the reliability of measures and the use of pretests (or covariates) both have the potential to increase the statistical power in research designs. Tables are presented that show how specific changes in these factors can dramatically reduce the sample size necessary to determine whether an intervention is effective, thus potentially improving the effectiveness of assessment in terms of reducing the time and cost of collecting scientific support for new practices.

Schools, colleges, and universities are increasingly turning to the assessment of learning outcomes to evaluate the effectiveness of their programs. Unfortunately, for institutions with few students per year, it may take years to accumulate a large enough sample size to form statistically sound conclusions about the effectiveness of an instructional practice. Even for institutions with larger student populations, the collection of a large sample may be cost-prohibitive. This paper shows how improving the reliability of outcome measures and including pretests or covariates in the assessment process increase statistical power and can dramatically reduce the required sample size, thus enabling an organization to collect supportive evidence more quickly and inexpensively. Some background on statistical power in research designs, and the use of pretests in particular, will be offered before presenting tables that describe the interrelationships among assessment reliability, pretests, power, and sample size.

Background

Statistical Power in Research Designs

Assessment measures are often used as part of a tracking study of some sort, and often in cohort designs. Assessment measures may include exams, such as a final given to all students who complete an algebra course. The same exam can be given each semester, and the data then combined to enable tracking of improvements in algebra achievement from semester to semester or year to year. In the cohort design, the people who completed the school's program in the past can be seen as a control group (e.g., last year's algebra students) and those who completed the program more recently, after some instructional change occurred in the program, can be seen as a treatment group (e.g., this year's algebra students). These groups may also be represented in an experimental design using concurrent programs, where a control group receives the conventional form of education and the treatment group receives some new form of instruction.

One of the keys to achieving meaningful results with experimental designs is statistical power, which is the ability to detect statistically significant differences. The greater the statistical power in an experiment, the greater the chances of finding a statistically significant result. One type of power analysis is the determination of how large a sample must be drawn in order to have a reasonable chance of achieving statistical significance when an effect is present. A commonly used threshold in these analyses is .80 (Cohen, 1977, p. 56), meaning the researcher asks how large a sample is necessary to have an 80% chance of detecting a statistically significant difference, if the expected difference exists.

The analysis of power is dependent on how large a difference the researcher expects to find. In most statistical analyses, it is assumed that the null hypothesis is true, that is, the control group and the treatment group have exactly the same outcomes. In power analyses, it is assumed that the control group and the treatment group will have some specific difference in outcomes. Assuming some specific difference, or effect size, is necessary to be able to estimate the probability of detecting a difference of that magnitude. Many methods are available for specifying the size of the expected difference in outcomes, but one common measure of effect size is the standardized mean difference in outcome measures (d), which Cohen (1977) describes as

(1) $$d = |m_1 - m_2| / s$$

In this equation, m_1 and m_2 are the mean scores from the two groups of students to be compared, and s is the standard deviation of the scores. Thus, d is essentially the mean difference between two groups in standard deviation units. For assessment professionals with some experience or expertise in an area, the *a priori* estimation of the expected differences

is not difficult. For example, Bloom (1976) offers estimates of effect sizes for a variety factors that affect educational outcomes, such as cognitive entry behaviors or classroom size.

In this paper, the analysis of power is developed from a test for differences across two groups, the *t*-test. Although the power of the *t*-test has been extensively studied and the importance of assessment reliability, the use of pretests, and effect size is well noted (see, for example, Lipsey, 1990), one unique aspect of this paper is that it shows how these factors affect sample size *in combination*. Starting with the basic equation for the power of a *t*-test, various substitutions are made to include the reliability of the assessment measures, the use of pretests, and the effect size (*d*) in the power equation (the detailed derivation is in the Appendix).

The Effects of Pretests and Covariates

The effect of a pretest on assessment effectiveness is primarily a function of how well that pretest correlates with the final assessment, or posttest. Higher pretest-posttest correlations indicate that the pretest explains more variance in the posttest, leaving less variance unexplained. Thus, if there is any variance in the final assessment due to the effect of a new instructional form, that variance will be more obvious. Although the use of meaningful pretests is generally possible, they are sometimes impractical. For example, in an educational setting, students may have little incentive to carefully complete a pretest on the first day of class, and so the pretest reliability may be poor. In addition, beginning students may have so little knowledge of a content area that their test responses are not consistent with a meaningful scale, and thus the pretest-posttest correlation may be so low that the pretest is virtually useless for statistical purposes. For example, Bacon (2002) reported pretest-posttest correlations from a junior-level business college course of around .30, even though the final exam had a reliability of .92. At this level, the pretest reduces the total variance in the posttest by only about 10% (see also Equation A6), an amount many would consider to be not worth the trouble (e.g., Reichardt, 1979).

When pretests are impractical, or when additional explained variance is desired, covariates may be considered. It is important to note that because the key characteristic of the pretest is its correlation with the posttest, any reasonable covariate or set of covariates may be used instead of or in addition to a pretest, as long as the experimental groups are equivalent. For example, in Bacon's (2002) study, although the pretest was not highly correlated with the final exam, grade point average was correlated with the final exam at the .63 level. Pretests and covariates alike are generally modeled as covariates in subsequent statistical tests (Reichardt, 1979). It should be noted, however, that in the case of non-equivalent group designs, using covariates might lead to bias if there are systematic differences across the groups (see especially Reichardt, 1979, p. 169). A pretest would be preferred under these conditions. For simplicity in describing the tables presented later in this paper, the variable or set of variables that may be used as covariates are lumped together and referred to as "the pretest."

It is also important to note that the observed pretest-posttest correlation ($\rho_{XYObserved}$) is related to the reliability of the posttest measure (ρ_{YY}) and the reliability of the pretest (ρ_{XX}) through the following the relation

(2) $$\rho_{XY(Observed)} = \sqrt{\rho_{XX}}\,\rho_{XY(True)}\sqrt{\rho_{YY}}\ .$$

Thus, in practice, improvements to the posttest measure reliability will also increase the pretest-posttest correlation.

Methodology

To understand the effects of assessment reliability, pretests, and effect size on study design, meaningful values for these variables will be inserted in the equations from the Appendix and the required sample sizes tabulated. Cohen (1977, p. 56) makes a compelling case that a meaningful value for the acceptable level of power would be .80, and recognizes .05 as a very commonly used α level, so these values will be used in the analysis here. In terms of reliability, Cohen (1977) notes that reliabilities (ρ_{yy}) of around .70 are commonly observed. Nunnally (1978, p. 245) considers .70 to be an acceptable level for early stages of basic research, but suggests that .80 or higher would be more appropriate in some applied settings. In pilot tests, especially with performance assessments (Sax, 1997, p. 167), lower reliabilities may often be observed. Therefore, the set of reliabilities of .5, .7, and .9 will be used here.

In considering meaningful values of the pretest-posttest relationship for the present analysis, one must recognize that in practice many studies use no pretest at all, effectively setting the pretest-posttest correlation (ρ_{XY}) equal to 0. At the high end of $\rho_{XY,}$ it should be noted that the correlation with the pretest cannot exceed the square root of the reliability of the posttest (Equation 2). Therefore, the set of pretest-posttest correlations of 0, .3, .5, and .7 will be used here.

Meaningful values for effect sizes can be found in Cohen (1977), although different standards may be appropriate for academic research in educational psychology (see Osborne, 2003). Cohen (1977) describes small, medium, and large standardized mean differences (*d*) in the social sciences as .2, .5, and .8, respectively. (These correspond to R^2 values of .01, .09, and .25 respectively, p. 79.) These guidelines reflect the assumption that the reliability of measurement in these studies is often in the neighborhood of .70 (p. 79-81). From Equation A5, we can see that when ρ_{yy} is .70 and σ^2_{True} is 1.0 (as standardized here), the total variance would be 1.43 (1.0/.70) and thus the observed standard deviation would

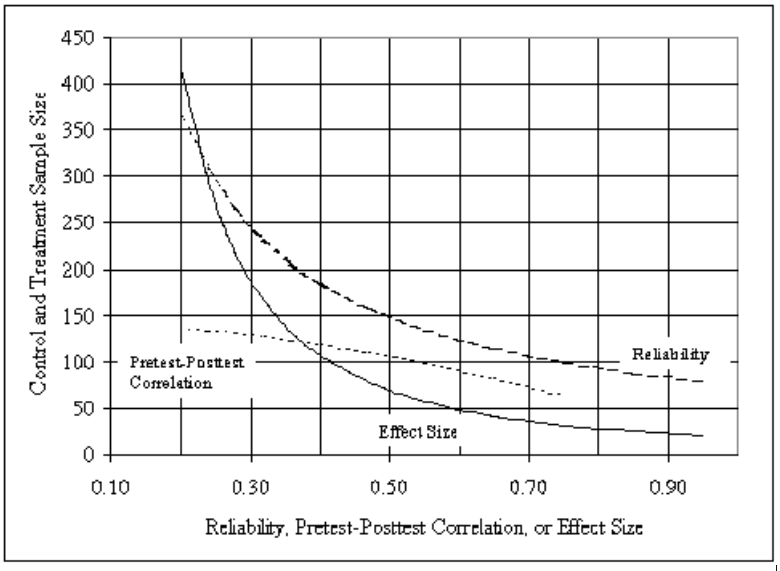
be 1.20. Therefore, to model Cohen's observed effect sizes of .2, .5, and .8 will require "true effect sizes" of .24, .60, and .96, respectively (.2x1.2, .5x1.2, and .8x1.2, respectively). These true effect sizes (ES_T) are used in the tables and the figure presented in this paper.

Setting the left side of Equation A1 at .80, and substituting the values mentioned above for ρ_{YY}, ρ_{XY}, and ES_T into the right side (via Equation A9), the sample size value that balanced the equation was noted. The GoalSeek function in MS Excel was used to solve for sample size values at various levels of the other variables in the model. In the special cases where $\rho_{yy} = .70$ and $\rho_{XY} = 0$, the results were identical (within rounding error) to those reported in Table 2.3 in Cohen (1977).

Results and Discussion

The effects of reliability, the use of pretests, and the effect size on the required sample sizes are shown in Figure 1. (The figure was simplified by setting the control and treatment group sizes equal, but later analyses will allow these to differ.) Figure 1 shows that over the range of values shown here, the effect size appears to have the most dramatic impact on the size of the sample required. An effect size (ES_T) of .2 would require 840 subjects (420 each in the control and treatment groups) to have an 80% chance of detecting a significant difference (at the .05 level, two-tailed). However, if the effect size were .95, the required total sample would be only 40 (20 each in each group, assuming the pretest-posttest correlation, ρ_{XY}, = .50 and the outcome measure reliability, ρ_{YY}, = .70). The reliability of the outcome measure and the use of a pretest were similar in their effectiveness over a reasonable range of values. The required group sizes would drop from about 150 to about 75 as the reliability increases from .5 to .95 (assuming $\rho_{XY} = .50$, $ES_T = .40$), while the required group sizes would drop from about 140 to about 60 as the pretest-posttest correlation increases from .2 to .75 (assuming $\rho_{YY} = .70$ and $ES_T = .40$). Although effect size is found to be a major driver of sample size, in practice an assessment professional may not have control over the size of the improvements in outcomes. Also, as noted earlier, in practice improvements in outcome measure reliability will generally lead to improvements in the pretest-posttest correlation. Therefore, we can conclude from this analysis that outcome measure reliability and the effective use of pretests both warrant close attention in the design of assessment plans.

Figure 1: The effects of reliability, pretest-posttest correlation, and effect size on the sample sizes necessary to have an 80% chance of detecting a statistically significant difference.



Note: Default parameters include alpha level = .05, pretest-final correlation = .50, reliability = .70, true effect size = .40.

Tables 1, 2, and 3 show the sample sizes required for small, medium, and large effect sizes, respectively, given various assessment reliabilities and pretest-posttest correlations. A hypothetical example will be offered to show how these tables could be used to evaluate an assessment system. Suppose a school decided to implement an assessment system wherein the learning outcomes of students would be assessed each year, using tests or similar measures, and these outcomes would be compared to prior years. The system would include statistical analyses to evaluate the effect of changes in the curriculum, and these analyses would form the basis of decisions to retain or reject changes. Thus, the system can be seen as a closed feedback loop. Suppose this system had been in place for one year, and at the end of that year, a new curriculum was to be introduced that was expected to lead to improvements in learning outcomes. The assessment system didn't use pretests and used a locally-developed outcome measure with a reliability of only .50. Such a system might be obtained from portfolio-based assessments with poorly trained graders or weak rubrics, or from tests that have not been subjected to item analysis. At this school, 60 students finish this level each year, so the control sample is 60. Suppose further that the changes to be expected in outcomes corresponded to medium effect sizes (based on studies reported in Bloom, 1976). Table 2 (second column from the left) shows that a sample of 166 more students

would be needed to comprise the treatment group in order to have an 80% chance of detecting a statistically significant improvement. At this school, the assessment system would take nearly three more years (nearly four all together, including the control year) to collect the data before a sufficiently powerful statistical evaluation could be made.

Table 1: Required treatment sample sizes for a given control sample (n_C), reliability (ρ_{YY}), and pretest-posttest correlation (ρ_{XY}) and a small true expected effect size ($ES_T = .24$).

	ρ_{YY} = .5				ρ_{YY} = .7				ρ_{YY} = .9			
n_C	ρ_{XY} = 0	ρ_{XY} = .3	ρ_{XY} = .5	ρ_{XY} = .7	ρ_{XY} = 0	ρ_{XY} = .3	ρ_{XY} = .5	ρ_{XY} = .7	ρ_{XY} = 0	ρ_{XY} = .3	ρ_{XY} = .5	ρ_{XY} = .7
10	*	*	*	*	*	*	*	*	*	*	*	*
20	*	*	*	*	*	*	*	*	*	*	*	*
40	*	*	*	*	*	*	*	*	*	*	*	*
60	*	*	*	*	*	*	*	*	*	*	*	*
100	*	*	*	*	*	*	*	10908	*	*	*	345
150	*	*	*	1904	*	*	5090	298	*	1708	473	162
200	*	*	*	460	7063	1564	542	199	629	447	266	127
400	855	656	420	214	381	319	231	133	244	211	159	96
1000	375	330	257	162	242	216	171	110	179	160	128	84
2000	315	283	228	149	216	194	158	104	164	148	120	81
1000000	273	248	204	139	195	177	146	99	151	138	113	77
$n_C = n_T$	546	497	410	279	390	355	293	200	304	276	228	156

* No n_T sample size would be sufficient to achieve 80% power under these conditions.

Table 2: Required treatment sample sizes for a given control sample (n_C), reliability (ρ_{YY}), and pretest-posttest correlation (ρ_{XY}) and a medium expected true effect size ($ES_T = .60$).

	ρ_{YY} = .5				ρ_{YY} = .7				ρ_{YY} = .9			
n_C	ρ_{XY} = 0	ρ_{XY} = .3	ρ_{XY} = .5	ρ_{XY} = .7	ρ_{XY} = 0	ρ_{XY} = .3	ρ_{XY} = .5	ρ_{XY} = .7	ρ_{XY} = 0	ρ_{XY} = .3	ρ_{XY} = .5	ρ_{XY} = .7
10	*	*	*	*	*	*	*	*	*	*	*	*

20	*	*	*	*	*	*	*	86	*	*	217	36
40	*	4060	189	53	148	103	60	28	65	52	35	19
60	166	122	75	37	67	56	40	22	42	36	27	16
100	79	67	50	29	46	40	31	19	33	29	23	14
150	62	55	42	27	40	35	28	18	29	26	21	14
200	57	50	40	25	37	33	27	17	28	25	20	13
400	49	44	36	24	34	31	25	17	26	23	19	13
1000	46	41	34	23	32	29	24	16	25	23	19	13
2000	45	41	33	22	32	29	24	16	25	22	18	12
1000000	44	40	33	22	31	28	23	16	24	22	18	12
$n_C = n_T$	88	81	67	46	63	58	48	33	50	45	38	26

* No n_T sample size would be sufficient to achieve 80% power under these conditions.

Table 3: Required treatment sample sizes for a given control sample (n_C), reliability (ρ_{YY}), and pretest-posttest correlation (ρ_{XY}) and a large expected true effect size ($ES_T = .96$).

	ρ_{YY} = .5				ρ_{YY} = .7				ρ_{YY} = .9			
n_C	ρ_{XY} = 0	ρ_{XY} = .3	ρ_{XY} = .5	ρ_{XY} = .7	ρ_{XY} = 0	ρ_{XY} = .3	ρ_{XY} = .5	ρ_{XY} = .7	ρ_{XY} = 0	ρ_{XY} = .3	ρ_{XY} = .5	ρ_{XY} = .7
10	*	*	*	80	*	*	128	20	213	77	31	12
20	128	78	40	17	35	28	19	10	20	17	12	7
40	31	27	20	12	19	16	13	8	13	12	9	6
60	25	22	17	11	16	14	11	7	12	10	8	5
100	21	19	15	10	14	13	10	7	11	10	8	5
150	20	18	14	9	13	12	10	7	10	9	8	5
200	19	17	14	9	13	12	10	6	10	9	7	5

400	18	16	13	9	13	11	9	6	10	9	7	5
1000	17	16	13	9	12	11	9	6	10	9	7	5
2000	17	16	13	9	12	11	9	6	10	9	7	5
1000000	17	16	13	9	12	11	9	6	9	9	7	5
$n_C = n_T$	35	32	27	19	25	23	19	14	20	18	15	11

* No n_T sample size would be sufficient to achieve 80% power under these conditions.

To continue the hypothetical example, now imagine that an assessment professional noticed the low reliability of the measures in a pilot study and, realizing that four years is too long to wait for feedback, immediately made changes to the measures. Suppose the outcome measure's reliability were improved to .90, and a pretest (perhaps prior grades, or standardized test scores from another source, if the treatment and control groups were fairly equivalent) were added that correlated with the outcome measure at the .70 level. Still looking in Table 2 (the underlying effect size has not changed), the far right column indicates that now only 16 students would be necessary for the treatment sample. Further, if the school was able to implement the control and the pretest in the same year, perhaps to different groups of students, the size of each group need only be 26 (from the bottom row of Table 2). Thus, the school's assessment system could make a sufficiently powerful quantitative evaluation in one year or less. Of course, with smaller sample sizes, the researcher must always watch for outliers, and statistical significance at the .05 level may not always be necessary to make reasonable decisions, but from this example it is clear that by improving the reliability of the measures and using predictive pretests, assessment systems may obtain sound conclusions in a much more timely manner.

Concluding Remarks

The analysis and the examples presented here demonstrate how improvements in reliability of outcome measures and the use of predictive pretests (or covariates if appropriate) can lead to striking improvements in assessment systems. By reducing the sample sizes required for sound assessment, assessment systems so improved may provide feedback to program administrators in a much more timely and cost-effective manner. The systematic collection and distribution of measures that may function as useful covariates may thus be seen as an important aspect of building a school's assessment system. Conversely, if reliability and pretest issues are ignored, assessment systems may amount to little more than bureaucratic overhead with little hope of providing useful information. Increased attention to the measurement issues described here may therefore be essential to the success of assessment programs.

Notes

The author gratefully acknowledges the insightful comments received from Kim A. Stewart, Charles S. Reichardt, Melvin M. Mark, and two anonymous reviewers on earlier drafts. The financial assistance from the Daniels College of Business that was instrumental in completing this research is also gratefully acknowledged.

Appendix

This appendix shows how the power of a t-test used to compare a control group with a treatment group in a pretest-posttest experimental design can be described in terms of the reliability of the posttest, the pretest-posttest correlation, the sample size, and the effect size. The development starts with an approximation of the power of the t-test, from Hays (1981, p. 288).

(A1) Power = 1 – prob. $\left\{ z \leq \left(\frac{t' - \delta}{\sqrt{1 + \frac{(t')^2}{2v}}} \right) \right\}$

where v is the degrees of freedom, or the size of the control group plus the size of the treatment group minus 2 ($n_C + n_T - 2$),

t' is the critical value of the t statistic evaluated at a selected α level with v degrees of freedom, and

δ is the noncentrality parameter.

The noncentrality parameter in Equation A1 captures the expected difference between the two groups, and is closely related to Cohen's (1977) *d*, described in the body of the paper. The noncentrality parameter is analyzed more closely here in order to model the factors that contribute to the standard deviation of the mean difference. In more detail, then, the noncentrality parameter is a function of the expected difference in the populations divided by the standard error of that difference, or

(A2) $\text{est.}\ \delta = \dfrac{est(|\mu_C - \mu_T|)}{est.\sigma_{diff.}}$

where $|\mu_C - \mu_T|$ represents the absolute difference in means between the control group and the treatment group and $\sigma_{diff.}$ is the estimate of the pooled standard error of the mean difference. For simplicity in this analysis, we make the common assumption that the variance in the two groups is the same, so $\sigma_{diff.}$ can be estimated (Hayes, 1981, p. 285) as

(A3) $\text{est.}\ \sigma_{diff.} = \sqrt{\sigma^2\left(\dfrac{1}{n_C} + \dfrac{1}{n_T}\right)}$,

This analysis is now extended to accommodate differences in the reliability of the outcome measure, ρ_{yy} (also assumed to be the same across treatment and historical groups). Note that reliability is the ratio of true score variance, σ^2_{True}, to total variance, σ^2_{Total} (Nunnally, 1978, p. 200), or

(A4) $\rho_{YY} = \dfrac{\sigma^2_{True}}{\sigma^2_{Total}}$.

This can be re-arranged to show how the total observed variance is a function of the true variance and the reliability of the measure, or

(A5) $\sigma^2_{Total} = \dfrac{\sigma^2_{True}}{\rho_{YY}}$.

From Equation 5 it is clear that if we assume that the variance in the underlying phenomenon (σ^2_{True}) remains the same no matter how reliably it is measured, then increasing the reliability will be equivalent to reducing the error variance, thus reducing the total variance and increasing the power of the test.

The total observed variance can also be reduced by the use of a pretest. The critical property of this pretest in this analysis is the correlation between the pretest and the final outcome assessment (the posttest), or ρ_{XY}. The remaining unexplained variance, σ^2_U, after controlling for the pretest would be (Reichardt, 1979, p. 157)

(A6) $\sigma^2_U = \sigma^2_{Total}(1 - \rho^2_{XY})$

From Equation A6 it should be clear that as the correlation between the pretest and posttest (ρ_{XY}) increases, the remaining unexplained variance decreases. The combined variance-reducing effects of outcome measure reliability and the use of pretests can be described by combining Equations A5 and A6, yielding

(A7) $\sigma^2_U = \dfrac{\sigma^2_{True}(1 - \rho^2_{XY})}{\rho_{YY}}$.

Substituting this more detailed analysis of the variance in outcome measures into Equation A3 yields

(A8) $\text{est.}\ \sigma_{diff.} = \sqrt{\dfrac{\sigma^2_{True}(1 - \rho^2_{XY})}{\rho_{YY}}\left(\dfrac{1}{n_C} + \dfrac{1}{n_T}\right)}$.

The relationships among the quantities of interest can now be generally described by assuming a type of standardization of the measures. The difference in the means, $|\mu_C - \mu_T|$, will now be described in terms of Cohen's standardized effect size, *d*. Cohen describes the effect size of a comparison of means as the difference in means divided by the standard deviation (Cohen, 1977, p. 20). To standardize the analysis used here, σ^2_{True} will be set to 1.0, creating a measure of the "true" standardized effect. Combining Equations A2 and A8 yields a more complete formula for the noncentrality parameter as

(A9) est. $$\delta = \frac{ES}{\sqrt{\frac{(1-\rho_{XY}^2)}{\rho_{YY}}\left(\frac{1}{n_C}+\frac{1}{n_T}\right)}}.$$

Substituting Equation A9 into Equation A1 yields a relation describing the power of a *t*-test that integrates the posttest reliability, the pretest-posttest correlation, the sample size, and the effect size.

References

Bacon, D.R. (2002, October). *Issues in the use of pretests*. Paper presented at the Colorado Regional Higher Educational Assessment Conference, Westminster, CO.

Bloom, B.S. (1976). *Human characteristics and school learning*. New York: McGraw-Hill.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences, revised edition*. New York: Academic Press.

Hays, W.L. (1981). *Statistics, 3rd edition*. New York: Holt, Rinehart and Winston.

Lipsey, M.W. (1990). *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage.

Nunnally, J.C. (1978). *Psychometric theory, 2nd edition*. New York: McGraw-Hill.

Osborne, J.E. (2003). Effect sizes and the disattentuation of correlation and regression coefficients: Lessons from educational psychology. *Practical Assessment, Research & Evaluation, 8*(11). Retrieved September 30, 2003 from http://edresearch.org/pare/getvn.asp?v=8&n=11.

Reichardt, C.S. (1979). The statistical analysis of data from nonequivalent group designs. In Cook, T.D., & Campbell, D.T., *Quasi-experimentation: Design & analysis issues for field settings,* pp. 147-205. Boston: Houghton Mifflin.

Sax, G. (1997). *Principles of educational and psychological measurement and evaluation, 4th edition*. Belmont, CA: Wadswort h.

Address all correspondence to:

Donald R. Bacon
Department of Marketing
Daniels College of Business
University of Denver
2101 S. University Blvd (Rm 495)
Denver, CO 80208

Voice: 303-871-2707
Fax: 303-871-2323
E-mail: dbacon@du.edu

Descriptors: Reliability; Pretest; Covariate; Power; Sample Size

Citation: Bacon, Donald (2004). The contributions of reliability and pretests to effective assessment. *Practical Assessment, Research & Evaluation*, 9(3). Available online: http://PAREonline.net/getvn.asp?v=9&n=3.