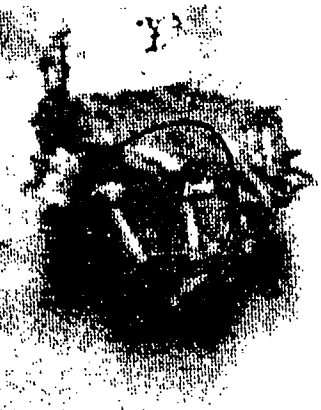

NAMRL SPECIAL REPORT 93-7

TEST RELIABILITY AND EXPERIMENTAL POWER

R.R. Stanny

Naval Aerospace Medical Research Laboratory
Naval Air Station
Pensacola, Florida 32508-5700

NAVAL AEROSPACE MEDICAL RESEARCH LABORATORY
51 HOVEY ROAD, PENSACOLA, FL 32508-1046

NAMRL SPECIAL REPORT 93-7

TEST RELIABILITY AND EXPERIMENTAL POWER

R.R. Stanny

Reviewed and approved 20 Dec 93

A. J. MATECZUN, CAPT, MC USN
Commanding Officer

This research was sponsored by the Walter Reed Army Institute of Research under Military Interdepartmental Purchase Request No. 90MM0523. The project was monitored by the Naval Medical Research and Development Command under work unit 637.64A3M4637648995.AB-088.

The views expressed in this article are those of the author and do not reflect the official policy or position of the Department of the Navy, Department of Defense, nor the U.S. Government.

Trade names of materials and/or products of commercial or nongovernment organizations are cited as needed for precision. These citations do not constitute official endorsement or approval of the use of such commercial materials and/or products.

ABSTRACT

The reliability coefficient, ρ_{XX}, has long been accepted as an index of the stability, repeatability, and precision of psychological tests. Because ρ_{XX} measures the proportion of the variance in a set of scores attributable to variation among individuals, values of ρ_{XX} are sometimes compared to justify using particular tests in studies of individual differences. Values of ρ_{XX} are also sometimes compared to justify using particular tests in experimental research. The latter practice is usually justified by arguing that larger values of ρ_{XX} imply greater measurement precision and, therefore, potentially greater sensitivity to experimental treatments. That argument is not generally correct because the individual variation measured by ρ_{XX} is frequently confounded with measurement error in the denominators of significance tests. The effects of this confounding lead to "paradoxical" situations in which reliability, as measured by ρ_{XX}, may be inversely related (or unrelated) to experimental precision, as measured by the reciprocal of experimental error. Because the power of an experiment increases with precision, as just defined, conditions that invert or negate the relationship between ρ_{XX} and precision also invert or negate the relationship between ρ_{XX} and power. These considerations do not mean that the reliability coefficient is necessarily irrelevant to experimental research. Because experimental designs differ in the degree to which they are influenced by individual variation, a consideration of the value of ρ_{XX} a specific test yields will sometimes provide information about the best design in which to use that test.

ACKNOWLEDGMENTS

I gratefully thank C. J. Stanny, J. A. D'Andrea, J. O. deLorge, C. E. Williams, D. Zimmerman, and K. S. Mayer for valuable critiques of earlier versions of this manuscript.

INTRODUCTION

The reliability coefficient, ρ_{XX}, plays a fundamental role in studies of individual differences because it expresses the proportion of variance in a set of scores attributable to variation among individuals (Gulliksen, 1950). Hence, ρ_{XX} is often described as an index of the precision or accuracy of tests (Kerlinger, 1986; Lord & Novick, 1968). For these reasons, values of ρ_{XX} are sometimes compared to justify decisions to use particular tests in studies of individual differences (Weiss & Davison, 1981).

Values of ρ_{XX} are also sometimes compared to justify a decision to use a particular psychological test in experimental studies. Common sense suggests that, if the reliability coefficient measures precision, a respectable value of ρ_{XX} should be necessary for a test to be sensitive to the effects of an experimentally manipulated independent variable. Variants of this proposition have been accepted by numerous authors (e.g., Carter, Krause, & Harbeson, 1986; Cleary & Linn, 1969; Cook & Campbell, 1979; Fleiss, 1976; Humphreys & Drasgow, 1989a, 1989b; NATO Aerospace Medical Panel Working Group 12, 1989; Sutcliffe, 1958). Unfortunately, a policy of using reliability coefficients to judge the relative sensitivities of different psychological tests can yield misleading results. Reasons why this is so are outlined in the remainder of this paper.

In the subsections that follow, I will outline the statistical issues as they pertain to simple between-groups and repeated-measures designs. In the Discussion, I will review the controversy surrounding the interpretation of the statistical results and examine a widely held informal argument according to which ρ_{XX} is directly related to power. We will see that in none of the experimental designs considered here should a comparison of the reliability coefficients of different psychological tests be expected to indicate which of several different tests is likely to be more sensitive to the effects of an experimental treatment. On the other hand, we will see that a knowledge of one test's reliability coefficient can sometimes help an investigator determine the experimental design in which that test will be most sensitive to the effects of an experimental treatment.

The relationship between ρ_{XX} and experimental power was addressed some years ago by Overall and Woodward (1975) who offered the "paradoxical" observation that, if measurement error is held constant, the power of an analysis of difference scores is maximized when the reliability coefficient of the differences is zero. The validity of their observation has been obscured by the contentious and occasionally confusing interchange that followed. To understand the psychometric basis of the argument, recall that in classical test theory an observed test score, X_i, is assumed to be the sum of a true score, T_i, and a measurement error, E_i (e.g., Gulliksen, 1950). Measurement errors are assumed to be random with a mean of zero, and to be independent of the true scores and of each other. Hence the variance of a set of test scores is a sum of true-score and measurement-error variances; i.e.,

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2, \tag{1}$$

where σ_X^2 is the variance of the observed scores, σ_T^2 the variance of the true scores, and σ_E^2 the variance of the measurement errors. The reliability coefficient, in turn, is the proportion of the scores' variance attributable to variance in true scores (e.g., Gulliksen, 1950). That is:

$$\rho_{XX} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{(\sigma_T^2 + \sigma_E^2)}. \tag{2}$$

Perhaps the most familiar estimate of ρ_{XX} is the test-retest correlation, which is determined by obtaining scores from the same individuals on two occasions and calculating the Pearson product-moment correlation between first and second scores.

BETWEEN-GROUPS CONTRASTS

First, consider a simple between-groups design in which a psychological test is administered to treatment and control groups to measure the effect of an experimentally manipulated independent variable. Under the null hypothesis that group means are equal, the independent-samples t test for groups of equal sizes can be written:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{[(\hat{\sigma}_{X1}^2 + \hat{\sigma}_{X2}^2) / n]^{1/2}}, \tag{3}$$

in which $\bar{X}_1$ and $\bar{X}_2$ are the group means, $\hat{\sigma}_{X1}^2$ and $\hat{\sigma}_{X2}^2$ are the estimated within group variances, and n is group size. Suppose, for simplicity, that true score and measurement error variances do not differ between groups. If we replace $\hat{\sigma}_X^2$ in Equation 3 with the right side of Equation 1, the equation for t becomes:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{[2\,(\hat{\sigma}_T^2 + \hat{\sigma}_E^2) / n]^{1/2}}. \tag{4}$$

Equation 4 indicates that, if the difference $\bar{X}_1 - \bar{X}_2$ is constant, the value of t (and, therefore, the sensitivity of the test) varies inversely with the summed magnitudes of the true and measurement-error variances. Furthermore, Equation 4 indicates that, in designs for which Equation 3 is appropriate, the value of Student's t (and the sensitivity of the test) will be unrelated to the *relative* magnitudes of true and measurement-error variances.[1] If the value of t is unrelated to the relative magnitudes of true and measurement-error variances in this design, t must also be unrelated to the value of the reliability coefficient in this design. This is because, as Equation 2 indicates, the reliability coefficient is determined by the relative magnitudes of true and error variances (Williams & Zimmerman, 1989; Zimmerman & Williams, 1986). An analogous derivation for the analysis of variance has been presented by Nicewander and Price (1978).[2]

To summarize: In between-groups experimental designs for which Equation 3 is appropriate, tests with equal total variances yield equal power; tests with different total variances yield different levels of power. Because

---

[1]One might argue that it would be more appropriate to phrase Equations 3-7 in effect-size notation rather than t-test notation. I have used t-test notation because to use effect-size notation would reduce clarity without affecting the conclusions.

[2]It is worth noting that Sutcliffe (1980) also presented an equivalent derivation for the analysis of variance, in an article occasionally cited as supporting the idea that experimental power necessarily increases with ρ_{XX}.

power varies with the sum of σ_T^2 and σ_E^2, not their relative magnitudes, power in between-groups designs is unrelated to ρ_{XX}.[3]

REPEATED-MEASURES CONTRASTS

Repeated-measures contrasts without subject-by-treatment interactions

Repeated-measures designs present somewhat different issues. Overall and Woodward (1975) considered the case of difference scores calculated by subtracting subjects' scores in one experimental condition from their scores in another. Difference scores of this type form the basic data of *t* tests for correlated observations. Overall and Woodward (1975) considered a model in which subjects do not differ in their responses to the experimental treatment (i.e., a model in which no subject-by-treatment interaction occurs). When the null hypothesis is that the average difference between scores in two experimental conditions is zero, the equation for the correlated *t* test can be written:

$$t = \frac{\bar{d}}{\hat{\sigma}_{\bar{d}}} \tag{5}$$

where $\bar{d}$ is the mean difference score and $\hat{\sigma}_{\bar{d}}$ is its estimated standard error. If individual differences are assumed equal in the two experimental conditions (the usual assumption), variance attributable to individuals disappears from the variance of the difference scores (Overall & Woodward, 1975). Hence, the variance of the differences is simply the summed measurement-error variance of the original scores; i.e., $\sigma_d = 2\sigma_E^2$, where σ_E^2 is the measurement error variance of the original scores. Thus, the *t* for correlated observations can be written:

$$t = \frac{\bar{d}}{(2\hat{\sigma}_E^2 / n)^{1/2}}. \tag{6}$$

An examination of Equation 6 indicates that if $\bar{d}$ and *n* are treated as constants, the magnitude of *t* and, therefore, the power of a test of difference scores depends only on the magnitude of $\hat{\sigma}_E^2$ (Overall & Woodward, 1975, 1976).

Therefore, if one's goal is to select the psychological test with the greatest power, and one has no reason to suppose that one of the tests under consideration will yield a larger average difference, one should select the test with the smallest value of σ_T^2. The relevance of ρ_{XX} to power in this example depends on the relationship between $\hat{\sigma}_E^2$ and ρ_{XX}. It follows from Equation 2 that this relationship is $\sigma_E^2 = (1 - \rho_{XX})\sigma_X^2$. Hence, the reliability coefficients of the original test scores are, indeed, relevant to power in tests based on difference scores (a point made by Overall

---

[3]Many authors have pointed out that it is possible to specify conditions under which differences in the relative sizes of σ_T^2 and σ_E^2 lead to systematic changes in both reliability and power. The most important example occurs when either σ_T^2 or σ_E^2 remains constant from one test to the next and the other varies. Comparing Equations 2 and 4, one can see that if σ_T^2 remains constant while σ_E^2 varies, ρ_{XX} will vary directly with *t* (assuming, of course, the numerator of Equation 4 remains constant). The opposite result is obtained if σ_E^2 remains constant while σ_T^2 varies. In this case, *t* varies *indirectly* with ρ_{XX}. Whether it is ever plausible to assume that either σ_E^2 or σ_T^2 will remain constant from one test to the next is an open question.

and Woodward, 1975; 1976). The relevance of ρ_{XX} to power in this case is that the additional power afforded by using a specific test in a repeated-measures design (rather than in a between-groups design) increases directly with the value of ρ_{XX}. This does *not* mean that comparing the reliability coefficients of two tests will indicate which test will yield the more powerful contrasts: Because σ_T^2 and σ_E^2 can both be expected to vary from one test to the next, there will ordinarily be no reason to suppose that the test with the largest value of $\sigma_T^2 / (\sigma_T^2 + \sigma_E^2)$ will have the smallest value of σ_E^2 (an exception to this generalization is outlined in the Discussion). Thus, simple comparisons of the reliability coefficients of two tests should not be expected to indicate which test is more powerful.

Overall and Woodward were primarily concerned with showing that difference scores, although frequently possessing low reliability coefficients, do not necessarily yield contrasts of low power. They drew the seemingly paradoxical conclusion that (when measurement error is held constant) "the value of the test statistic is maximized when the reliability of the difference scores is zero" (Overall and Woodward, 1975, p. 86). To understand this conclusion, note that the reliability and total variance of a set of difference scores will both increase if variance attributable to individual differences is added to the error variance of the difference scores. However, the resulting increase in total variance would inflate the denominator of Equation 6, thereby reducing t and experimental power. Of course, the conclusion that reliability varies inversely with power when σ_E^2 is held constant does *not* mean that tests that yield unreliable difference scores will necessarily yield more powerful contrasts than tests that yield reliable difference scores (Overall and Woodward never implied that this would be true). This is because σ_E^2 can be expected to vary from one psychological test to the next.

Repeated-measures with subject-by-treatment interactions

Fleiss (1976) argued that the analysis of Overall and Woodward (1975) was based on the unrealistic assumption that individuals do not vary in their responses to independent variables. Fleiss considered an alternative repeated-measures model in which subjects may differ in their responses to the independent variable. When subjects differ in the way they respond to an independent variable, the variances of difference scores become $2\sigma_E^2 + 4\sigma_{\alpha s}^2$, rather than $2\sigma_E^2$ as in Equation 6 (Fleiss, 1976; Sutcliffe, 1980). The new term, $4\sigma_{\alpha s}^2$, represents variance attributable to a subject-by-treatment interaction.[4] Hence, the equation for the correlated t test in the presence of a such an interaction might be rewritten:

$$t = \frac{\bar{d}}{[(2\theta_E^2 + 4\theta_{\alpha s}^2) / n]^{1/2}} \qquad (7)$$

Fleiss argued that when the subject-by-treatment interaction variance is held constant, power is maximized when the reliability coefficient of the difference scores, ρ_{dd}, is maximized, not when $\rho_{dd} = 0$. The reliability coefficient of the difference scores can be understood to be the proportion of the total variance of the differences attributable to a subject-by-treatment interaction (Fleiss, 1976; Sutcliffe, 1980). Examining Equation 7, one can see that reducing measurement error while holding the interaction variance constant will cause t to increase. The reliability of the differences will also increase because the reduction in measurement error will reduce the total variance of the differences, thereby increasing the proportion of the variance attributable to the interaction. Hence, power will increase directly with the reliability of the difference scores when the *subject-by-treatment interaction* is held constant and measurement error is allowed to vary.

---

[4]Fleiss (1976) and Sutcliffe (1980) defined $\sigma_{\alpha s}^2$ as the within-cell interaction variance. Nicewander and Price (1983) have pointed out that, in discussions of ANOVA, $\sigma_{\alpha s}^2$ is conventionally used to refer to sums of within-cell interaction variances (e.g., Scheffé, 1959), in which case the variance of the differences becomes $2(\sigma_{\alpha s}^2 + \sigma_E^2)$. I will follow Fleiss and Sutcliffe's usage, here, for consistency with the argument at hand.

Overall and Woodward (1976) acknowledged Fleiss's point but noted that Fleiss had failed to address theirs. Their point had been that the reduction in reliability that occurs when constant individual differences are removed by calculating difference scores does not imply that contrasts based on difference scores must have low power. It was logical for Overall and Woodward to approach this issue by considering the effects of reducing σ_T^2 when σ_E^2 is held constant because the process of calculating difference scores eliminates constant individual differences that contribute to σ_T^2 without affecting the random errors that produce σ_E^2.

Thus, in the case of repeated-measures with subject-by-treatment interactions, if one's goal is to select the test with the greatest power, and there is no reason to suppose that one of the tests under consideration will yield a larger mean difference between conditions, the best strategy is to select the test with the smallest value of $2\sigma_E^2 + 4\sigma_{\alpha\pi}^2$. However, the process of estimating the subject-by-treatment interaction would require one to obtain data in the experimental conditions of interest. With that information in hand, one could simply calculate values of t (or, better yet, effect-size estimates) for each candidate test and use these values to compare the tests' sensitivities directly.

DISCUSSION

The considerations just outlined suggest that in practical situations an investigator should not expect that a simple comparison of the reliability coefficients of two different psychological tests will reveal whether one of the tests is likely to yield more powerful or precise measurements of the effects of an experimentally manipulated independent variable. Except in special cases, the reliability coefficients of different tests need not be directly related to the magnitudes of error terms derived from scores on those tests (Nicewander & Price, 1978, 1983; Sutcliffe, 1980; Williams & Zimmerman, 1989).

An important special case in which power and precision *are* directly related to ρ_{XX} occurs when an investigator compares two tests that produce identical true scores but different measurement error variances (Nicewander & Price, 1978). If two tests produce the same true scores, the tests will also produce the same values of σ_T^2. Hence, the test with the smaller value of σ_E^2 will be more reliable and more powerful (consider Equations 2 and 4). A state of affairs like this can occur in practice when one succeeds in increasing ρ_{XX} by increasing test length. An increase in test length will sometimes reduce the influence of measurement error, thereby increasing both ρ_{XX} and power. Nicewander and Price (1983) suggest that the familiar practice of increasing test length to increase reliability and power may be the source of the belief that greater reliability is always associated with greater power; Nicewander and Price, however, also present a numerical counterexample in which an increase in test length brought about by adding blocks of slightly nonparallel trials increases ρ_{XX} but *reduces* power.

Despite the straightforward nature of these results, the relationship between ρ_{XX} and experimental power has remained controversial. Some investigators, without contesting the statistical results, have objected to the broad conclusion (as it has sometimes been phrased) that reliability and power are unrelated in experimental studies (e.g., Humphreys & Drasgow, 1989a, 1989b; Sutcliffe, 1980). This objection is compelling because the reliability coefficient of a test is, in fact, relevant to the issue of which experimental design will afford the most powerful contrasts of scores from that specific test. For example, if an experimental treatment simply adds a constant to each score, the power of a within subjects design will exceed the power of a between subjects design by an amount that increases directly with ρ_{XX} (recall the discussion of Equation 6).

A related objection derives from the philosophical idea that "reliability of measurement," properly defined, *should* be directly related to the power of experiments. Nicewander and Price (1983) have noted that portions of Sutcliffe's (1980) argument that reliability and power are directly related appear to imply that reliability might be more appropriately defined as a function of the reciprocal of measurement error. Nicewander and Price (1983) suggested that such a redefinition would solve the problem for contrasts based on difference scores. However, consideration of Equations 4 and 7 indicates that defining "reliability" as a function of $1 / \sigma_E^2$ would not necessarily

make experimental power a direct function of "reliability" when σ_T^2 or σ_{a*}^2 differ from one test to the next. Humphreys and Drasgow (1989a, 1989b), in contrast, have suggested that it would be possible to ensure that the reliabilities of difference scores always vary directly with power by incorporating the magnitudes of treatment effects into the definition of the coefficient. However, as Overall (1989) noted in a response to Humphreys and Drasgow, redefining the reliability coefficient as an index of effect size would require abandoning the long-standing tradition of interpreting the reliability coefficient as a measure of sensitivity to individual differences.

An informal argument sometimes offered to support the idea that larger values of ρ_{XX} tend to be associated with greater power is based on the notion that tests with relatively high values of ρ_{XX} are relatively sensitive to individual variation in true scores. According to this argument, if a test is relatively sensitive to variation in true scores, it should also be relatively sensitive to experimentally induced changes in true scores. This logic is correct if an additional assumption is valid. The additional assumption is that true scores on the tests being compared bear equivalent functional relationships to the same set of underlying psychological variables. When this assumption is valid, values of σ_T^2 obtained from all tests should be equal, and the average effect of experimental treatments on true scores should also be equal. When these conditions hold, the test with the largest reliability coefficient will be the test with the smallest σ_E^2. Assuming that all other factors are equal (as they should be under these assumptions), the test with the smallest value of σ_E^2 will yield the most powerful contrasts (see Equations 4, 6, and 7).

Some hazards of applying the logic just described to tests that are not essentially variants of a single test follow from possibilities that: (1) the underlying psychological processes that determine the true components of test scores may differ from one test to the next and/or (2) the functions that relate true scores to underlying processes may differ from one test to the next. To illustrate how such comparisons can go awry, suppose that tests A and B have equal measurement error but that true scores on A are determined by psychological attributes, Y1 and Y2, whereas true scores on B are determined only by Y1. If so,

$$\rho_{XX}(A) = \frac{\sigma_{Y1}^2 + \sigma_{Y2}^2}{\sigma_{Y1}^2 + \sigma_{Y2}^2 + \sigma_E^2}, \tag{9}$$

whereas

$$\rho_{XX}(B) = \frac{\sigma_{Y1}^2}{\sigma_{Y1}^2 + \sigma_E^2}, \tag{10}$$

where $\rho_{XX}(A)$ is the reliability of test A, $\rho_{XX}(B)$ is the reliability of test B, σ_{Y1}^2 is true variance due to attribute 1, σ_{Y2}^2 is true variance due to attribute 2, and σ_E^2 is measurement error. Because measurement-error variances are equal, Test A is more reliable than test B because its scores contain more true variance relative to error. (The additional true variance in scores from A is the variance attributable to Y2.) Although A is more reliable than B, the tests' relative sensitivities to experimental treatments will be impossible to predict if one does not know the causal structures that generate their scores.

For example, Test A will obviously be more sensitive than test B to treatments that affect only Y2 because test B is unaffected by changes in Y2. In between-groups designs, however, the less "reliable" test B will be more sensitive than test A to treatments that affect only Y1. This is because the error terms of between-groups significance tests derived from scores on test B will equal $[2(\hat{\sigma}_{Y1}^2 + \hat{\sigma}_E^2) / n]^{1/2}$, whereas those derived from test A will equal $[2(\hat{\sigma}_{Y1}^2 + \hat{\sigma}_{Y2}^2 + \hat{\sigma}_E^2) / n]^{1/2}$ (cf. Equation 4). Therefore, the error terms of between-groups significance tests derived from scores from test A will be inflated by irrelevant true variation in Y2. On the other hand, tests A

and B will yield equally powerful contrasts of treatments that affect only Y1 when they are used in within subjects designs. This is because the two tests yield equal within subjects error terms, $[2(\sigma_{\varepsilon}^2) / n]^{1/2}$ (see Equation 6).

As these examples illustrate, when it is incorrect to assume that tests being compared bear equivalent functional relationships to the same set of underlying psychological variables, ρ_{XX} and power may be directly related, inversely related, or unrelated. The direction and form of the relationship in any particular experiment will depend on the design of the experiment and the nature of the causal structure linking the independent and dependent variables.

Although comparing values of ρ_{XX} to justify the use of particular tests in experimental studies is hazardous, a test's reliability coefficient can still be relevant to the power of the experimental design in which the test is used. For example, the power of a within subjects contrast, relative to that of a between subjects contrast, varies directly with the correlation between subjects' scores in the different treatment conditions. As Overall and Woodward (1975) have pointed out, this correlation equals ρ_{XX} if the independent variable simply adds a constant to each score (i.e., if subjects and treatments do not interact). Furthermore, knowledge of ρ_{XX} can be sometimes be used to judge whether power can be increased more efficiently by adding a pretest and using its scores as covariates or by increasing the length of the posttest (Maxwell, Cole, Arvey, & Salas, 1991). Moreover, knowledge of the prepost correlation (which equals ρ_{XX} if subjects and treatments do not interact) can be used to judge whether a between-groups contrast of pretest-posttest difference scores will be more or less powerful than a simple contrast of posttest scores (Humphreys & Drasgow, 1989a; Kraemer & Thiemann, 1987; Overall & Ashby, 1991).

RECOMMENDATIONS

1. The reliability coefficient of psychometric theory, ρ_{XX}, should not be used as a surrogate effect-size estimate when selecting a test for use in a true experiment. This recommendation does not apply to (possibly rare) comparisons among tests that differ only in measurement error (σ_{ε}^2, as defined in psychometric theory), nor does it apply to nonexperimental research in which individual variation is a focus of interest.

2. The reliability coefficient of a test can sometimes be used to select the most powerful experimental design for use with that test. Hence, general statements to the effect ρ_{XX} is irrelevant to experimental research are incorrect.

REFERENCES

Carter, R. C., Krause, M., & Harbeson, M. M. (1986). Beware the reliability of slope scores. *Human Factors, 28*, 674-683.

Cleary, T. A., & Linn, R. L. (1969). Error of measurement and the power of a statistical test. *British Journal of Statistical and Mathematical Psychology, 22*, 49-55.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago. Rand McNally College Publishing Co.

Fleiss, J. L. (1976). Comment on Overall and Woodward's asserted paradox concerning the measurement of change. *Psychological Bulletin, 83*, 774-775.

Gulliksen, H. (1950). *Theory of mental tests.* New York: John Wiley & Sons.

Humphreys, L. G., & Drasgow, F. (1989a). Some comments on the relation between reliability and statistical power. *Applied Psychological Measurement, 13*, 419-425.

Humphreys, L. G., & Drasgow, F. (1989b). Paradoxes, contradictions, and illusions. *Applied Psychological Measurement, 13*, 429-431.

Kerlinger, F. N. (1986). *Foundations of behavioral research.* New York: Holt Rinehart and Winston.

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Maxwell, S. E., Cole, D. A., Arvey, R. D., & Salas, E. (1991). A comparison of methods for increasing power in randomized between-subjects designs. *Psychological Bulletin, 110*, 328-337.

NATO Aerospace Medical Panel Working Group 12. (1989, May). *AMP Working Group 12 and AGARD lecture series 163, AGARDograph No. 308: Human performance assessment methods.* Neuilly-Sur-Seine, France: North Atlantic Treaty Organization Advisory Group for Aerospace Research and Development.

Nicewander, W. A., & Price, J. M. (1978). Dependent variable reliability and the power of significance tests. *Psychological Bulletin, 85*, 405-409.

Nicewander, W. A., & Price, J. M. (1983). Reliability of measurement and the power of statistical tests. Some new results. *Psychological Bulletin, 94*, 524-533.

Overall, J. E. (1989). Distinguishing between measurements and dependent variables. *Applied Psychological Measurement, 13*, 432-433.

Overall, J. E., & Ashby, B. (1991). Baseline corrections in experimental and quasi-experimental clinical trials. *Neuropsychopharmacology, 4*, 273-281.

Overall, J. E., & Woodward, J. A. (1975). Unreliability of difference scores: A paradox for measurement of change. *Psychological Bulletin, 82*, 85-86.

Overall, J. E., & Woodward, J. A. (1976). Reassertion of the paradoxical power of tests of significance based on unreliable difference scores. *Psychological Bulletin*, *83*, 776-777.

Scheffé, H. (1959). *The analysis of variance*. New York: Wiley.

Sutcliffe, J. P. (1958). Error of measurement and the sensitivity of a test of significance. *Psychometrika*, *23*, 9-17.

Sutcliffe, J. P. (1980). On the relationship of reliability to statistical power. *Psychological Bulletin*, *88*, 509-515.

Weiss, D. J., & Davison, M. L. (1981). Test theory and methods. *Annual Review of Psychology*, *32*, 629-658.

Williams, R. H., & Zimmerman, D. W. (1989). Statistical power analysis and reliability of measurement. *Journal of General Psychology*, *116*, 359-369.

Zimmerman, D. W., & Williams, R. H. (1986). Note on the reliability of experimental measures and the power of significance tests. *Psychological Bulletin*, *100*, 123-124.

Other Related NAMRL Publications

None.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	December 1993	

4. TITLE AND SUBTITLE
Test Reliabiility and Experimental Power

5. FUNDING NUMBERS
C 90MM0523 WU 63764A
3M4637648995.AB-088

6. AUTHOR(S)
R.R. Stanny

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
NAVAEROMEDRSCHLAB
51 Hovey Road
Pensacola, FL 32508-1046

8. PERFORMING ORGANIZATION REPORT NUMBER
NAMRL Special Report
93-7

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Naval Medical Research and Development Command
National Naval Medical Center
8901 Wisconsin Avenue
Bethesda, MD 20889-5606

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The reliability coefficient, ρ_{XX}, has long been accepted as an index of the stability, repeatability, and precision of psychological tests. Because ρ_{XX} measures the proportion of the variance in a set of scores attributable to variation among individuals, values of ρ_{XX} are sometimes compared to justify using particular tests in studies of individual differences. Values of ρ_{XX} are also sometimes compared to justify using particular tests in experimental research. The latter practice is usually justified by arguing that larger values of ρ_{XX} imply greater measurement precision and, therefore, potentially greater sensitivity to experimental treatments. That argument is not generally correct because the individual variation measured by ρ_{XX} is frequently confounded with measurement error in the denominators of significance tests. The effects of this confounding lead to "paradoxical" situations in which reliability, as measured by ρ_{XX}, may be inversely related (or unrelated) to experimental precision, as measured by the reciprocal of experimental error. Because the power of an experiment increases with precision, as just defined, conditions that invert or negate the relationship between ρ_{XX} and precision also invert or negate the relationship between ρ_{XX} and power. These considerations do not mean that the reliability coefficient is necessarily irrelevant to experimental research. Because experimental designs differ in the degree to which they are influenced by individual variation, a consideration of the value of ρ_{XX} a specific test yields will sometimes provide information about the best design in which to use that test.

14. SUBJECT TERMS
Subject Terms: Statistics, psychometrics, reliability, power.

15. NUMBER OF PAGES
15

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	SAR