

Note on the Reliability of Experimental Measures and the Power of Significance Tests

Donald W. Zimmerman
Carleton University,
Ottawa, Ontario, Canada

Richard H. Williams
University of Miami,
Coral Gables, Florida

The statistical theory of the power of significance tests, combined with the classical theory of the reliability of measurement, reveals that the power of a statistical test sometimes increases and sometimes decreases as the reliability coefficient of a dependent variable increases. A seeming paradox that has been discussed extensively arises because the relation between statistical power and the reliability coefficient is not a functional relation unless another variable—either true variance or error variance—remains constant. This fact explains why authors have reached different conclusions about how reliability influences significance tests.

The relation between the power of statistical tests and the reliability of the experimental measures on which the tests are performed has been a subject of some concern and controversy in recent years. Several authors have proposed the idea that the power of a significance test can increase as the reliability of a dependent variable decreases. This apparent paradox was noted by Overall and Woodward (1975, 1976) and discussed extensively by Nicewander and Price (1978, 1983). (See also Cleary & Linn, 1969; Fleiss, 1976; Sutcliffe, 1958, 1980; Williams & Zimmerman, 1981.)

The relation between statistical power and reliability is important because, in the past, these two topics have been separate, and many applied statisticians and researchers are unconcerned with measurement theory and reliability. In this article we shall see that the paradox disappears if widely accepted, elementary results in statistical theory and measurement theory are considered together. This approach shows why some of the authors just mentioned have reached different conclusions as to how the reliability of a dependent variable influences the power of a significance test.

Relation Between Power and Reliability

It is well known that the power of a significance test, which by definition is the probability of detecting an experimental effect when one exists (or 1 minus the probability of a Type II error) is inversely related to the population variability of a dependent variable (see, e.g., Winer, 1971). Statistics texts commonly present families of power functions, or power curves as they are sometimes called, to illustrate how power as so defined depends on the variance of measures.

As far as the probability of rejecting the null hypothesis is concerned, it is immaterial whether this variability in measures

arises from differences in the true scores (group heterogeneity) or from the presence of error in the experimental measure (error of measurement). The greater the observed variability of a dependent variable, whatever its source, the less is the power of a statistical test designed to detect differences.

In classical measurement theory, the reliability of a measure is by definition the proportion of observed variance that is true variance or, equivalently, 1 minus the proportion of observed variance that is error variance. That is, $\text{Var } X = \text{Var } T + \text{Var } E$ and $\rho_{XX} = \text{Var } T / \text{Var } X = 1 - (\text{Var } E / \text{Var } X)$; where $\text{Var } X$ is observed variance, $\text{Var } T$ is true variance, $\text{Var } E$ is error variance, and ρ_{XX} is the reliability coefficient (Gulliksen, 1950; Lord & Novick, 1968). Furthermore, we know that the power of a significance test is a function of observed variance, $P = f(\text{Var } X)$, where f is a decreasing function.

If the observed variance of an experimental measure increases, one can be assured that the power of a statistical test using that measure will decrease. Yet without further information it is not possible to say whether such an increase in observed variance should be attributed to increasing true variance, increasing error variance, or both. In the first case, reliability increases along with variability; in the second case, reliability decreases as variability increases.

The power of a statistical test depends inversely on the magnitude of observed variance; in turn, observed variance is partitioned into true variance and error variance; and, finally, the reliability coefficient depends on the magnitude of the true and error components relative to each other. How, then, is power related to reliability?

In order to answer, it is necessary to be clear as to exactly what is meant by the question. A moment's reflection suggests that the meaning is as follows. Suppose the true variance of an experimental measure is fixed and that the reliability coefficient of the measure is increased by reducing error—that is by decreasing error variance. Then, how is the power of a statistical test that incorporates the measure influenced by such a change in reliability? The answer, unquestionably, is that power *increases* as reliability *increases*, because the increase in the reliability coefficient is accompanied by a decrease in observed variance.

Correspondence concerning this article should be addressed to Donald W. Zimmerman, Department of Psychology, Faculty of Social Sciences, Loeb Lab, Carleton University, Ottawa, Ontario, Canada K1S 5B6.

This reasoning appears to be straightforward. One cause of confusion is the fact that classical test and measurement theory suggests a different, although related question. It arises in the context of the topic of test reliability and group homogeneity (Gulliksen, 1950, p. 108; Lord & Novick, 1968, p. 129). Suppose the error variance in an experimental measure is fixed and that the reliability of the measure is increased by increasing the true variance—for example, by measuring a more heterogeneous group of subjects or examinees. Then how is the power of a statistical test influenced? The answer is that power *decreases* as the reliability coefficient *increases*, because the increase in reliability is accompanied by an increase in observed variance.

Although the somewhat counterintuitive answer to this second question does follow from well-known results in measurement theory, the question itself does not arise naturally in the statistical theory of hypothesis testing. Moreover, the latter theory has not been concerned explicitly with measurement or with the reliability coefficient, defined as a ratio of true variance and observed variance. These facts perhaps account to a large extent for the seeming "paradox" mentioned previously. It is conceivable that our question might be interpreted in other ways: for example, reliability could be increased or decreased by adjusting the magnitude of true variance and error variance in such a way that observed variance remains constant. Under these conditions, statistical power remains constant as reliability changes.

Power as a Function of Reliability

The question we raised earlier can be answered more formally as follows. First, statistical power is a monotonic decreasing function of observed variance—provided, of course, that the effect size, the sample size, and the significance level are held constant. Next, there is a *relation* between statistical power and reliability of measurement. However, it is not a *functional relation* unless one other variable—either true variance or error variance—has a definite value. Then, the correspondence between the reliability coefficient and statistical power is single-valued and one-to-one. Specifically, if the measurement-theoretic true variance is assumed to be constant, then power is a monotonic increasing function of the reliability coefficient. In addition, if the error variance is assumed to be constant, then power is a monotonic decreasing function of the reliability coefficient.

The different conclusions reached by the authors cited in the introduction can be explained by the fact that they asked different questions. Controversy might have been avoided if Overall and Woodward (1975) had phrased their question in more general terms: "Does statistical power *ever* increase as reliability decreases?" The answer is "Yes, whenever the decrease in reliability is accompanied by a decrease in the observed variance."

This happens whenever the ratio $\text{Var } T/\text{Var } X$ decreases at the same time that $\text{Var } X$ decreases—in other words, whenever a decrease in $\text{Var } X$ results from a decrease in $\text{Var } T$ that is proportionally greater than any decrease in $\text{Var } E$. This can happen in

the cases of simple difference scores considered by Overall and Woodward, although their ad absurdum example of $\text{Var } T = 0$ was neither necessary to make the point nor representative of cases that arise in practice.

The converse of the above question is "Does statistical power ever decrease as reliability increases?" Again, the answer is "Yes, whenever the increase in reliability is accompanied by an increase in observed variance." Several conditions described by Nicewander and Price (1978, 1983) can be accounted for in this way.

On the other hand, if one poses the question "Does statistical power always decrease as reliability increases?" the answer is certainly "No." On the contrary, statistical power increases as reliability increases whenever the ratio $\text{Var } T/\text{Var } X$ increases at the same time that $\text{Var } X$ decreases. This happens whenever a change in the ratio is produced by a decrease in $\text{Var } E$ that is proportionally greater than any decrease in $\text{Var } T$.

The latter interpretation is most pertinent to the concerns of researchers and applied statisticians. In experimental contexts, improvement in the reliability of a measure is usually interpreted as a reduction in error variance attributable to increased precision of an instrument or elimination of extraneous variables. Improvement in reliability in these contexts is not usually conceptualized as an increase in the heterogeneity of the group of subjects measured.

References

- Cleary, T. A., & Linn, R. L. (1969). Error of measurement and the power of a statistical test. *British Journal of Mathematical and Statistical Psychology*, 22, 49–55.
- 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: Wiley.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Nicewander, W. A., & Price, J. M. (1978). Dependent variable reliability and the power of statistical tests. *Psychological Bulletin*, 85, 405–409.
- Nicewander, W. A., & Price, J. M. (1983). Reliability of measurement and the power of statistical tests. *Psychological Bulletin*, 94, 524–533.
- 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.
- 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.
- Williams, R. H., & Zimmerman, D. W. (1981). Error of measurement and statistical inference: Some anomalies. *Journal of Experimental Education*, 49, 71–73.
- Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.

Received July 24, 1985

Revision received November 15, 1985 ■