Dependent Variable Reliability and Determination of Sample Size

Scott E. Maxwell
University of Houston

Arguments have recently been put forth that standard textbook procedures for determining the sample size necessary to achieve a certain level of power in a completely randomized design are incorrect when the dependent variable is fallible. In fact, however, there are several correct procedures—one of which is the standard textbook approach—because there are several ways of defining the magnitude of group differences. The standard formula is appropriate when group differences are defined relative to the within-group standard deviation of observed scores. Advantages and disadvantages of the various approaches are discussed.

One of the most frequently asked statistical questions in the social sciences is "How large must my sample be?" Implicit in this question is the recognition that the power of a statistical test must be considered, as well as the probability of a Type I error. The concept of power has received increasing attention in the social sciences in recent years (Brewer, 1972; Chase & Chase, 1976; Cohen, 1977; Schmidt, Hunter & Urry, 1976). Standard textbooks such as Glass and Stanley (1970), Hays (1973), Kirk (1968), and Winer (1971) contain extensive discussions of how to determine the appropriate sample size for a given experimental design.

In a series of articles, Levin and Subkoviak (1977, 1978) and Subkoviak and Levin (1977) have argued that the standard procedures advocated in textbooks are usually erroneous, even for a design as straightforward as a fixed effects completely randomized design. The basis for their argument is that earlier work (e.g., Cleary & Linn, 1969; Cleary, Linn, & Walster, 1970; Sutcliffe, 1958) showed that a decrease of error of measurement in the dependent variable is associated with increased statistical power; yet traditional formulas for determining sample size completely ignore error of measurement, despite the fact that almost all measures in the social sciences are fallible (i.e., they are not completely reliable). According to Subkoviak and Levin (1977), traditional textbook formulas will underestimate the actual sample size necessary to obtain a certain level of power, with the degree of underestimation directly related to the unreliability of the dependent measure. Hence, researchers following these formulas will take samples that are too small and will be too unlikely to reject the null hypothesis, when a reasonable alternative hypothesis is actually true.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 4, No. 2 Spring 1980 pp. 253–260

Traditional textbook formulas for determining power utilize an expression of the form

$$\phi = \sqrt{(n\psi^2)/(\nu+1)\ (\Sigma a_k^2)}\ , \qquad [1]$$

where

ϕ is a noncentrality parameter,
n is the sample size of each group,
ν is numerator degrees of freedom (K-1 in the one-way completely randomized design), and
ψ represents a standardized linear combination of interest, i.e.,

$$\psi = (\Sigma a_k \mu_k)/\sigma\ . \qquad [2]$$

Actually, this notation corresponds to that used by Subkoviak and Levin and differs somewhat from the notation used in most experimental design texts, where ϕ is defined as

$$\phi = \sqrt{n\Sigma(\mu_k-\mu)^2/(\nu+1)\sigma^2} \qquad [3]$$

when equal sample sizes are assumed. Equation 1 was developed by Levin (1975) to provide an expression for the power of a particular contrast, ψ, rather than for the omnibus test of no group differences, so that ϕ in Equation 1 must be less than or equal to ϕ in Equation 3. Equality is obtained by substituting

$$a_k = \mu_k - \mu \qquad [4]$$

for a_k in Equations 1 and 2. With this substitution,

$$\psi^2/\Sigma a_k^2 = \Sigma(\mu_k-\mu)^2/\sigma^2\ , \qquad [5]$$

so that with appropriate degrees of freedom, Equation 1 can also be used to calculate the power of the omnibus test. For this reason, Levin and Subkoviak's notation will be used here, with the understanding that the arguments also apply to Equation 3.

Reliability-Adjusted Formulas for Power

Subkoviak and Levin maintain that for fallible measures it is necessary to revise Equation 1 (and, implicitly, Equation 3) to take into account the degree of unreliability of the dependent measure. They argue that the appropriate formula for the noncentrality parameter in this case is

$$\phi = \sqrt{(\rho n\psi^2)/(\nu+1)\ (\Sigma a_k^2)}\ , \qquad [6]$$

where ρ is the reliability of the dependent variable. The inclusion of ρ in their formula is the only apparent modification of the traditional formula. In fact, however, as Forsyth (1978) has pointed out, there is an additional difference that proves to be very important. The traditional definition of ψ (labeled ψ_x for future reference) is

$$\psi_X = (\Sigma a_k \mu_k)/\sigma(X)\ , \qquad [7]$$

where $\sigma(X)$ is the within-group standard deviation of observed scores of the dependent measure. Subkoviak and Levin define ψ as

$$\psi_T = (\Sigma a_k \mu_k)/\sigma(T) \quad , \qquad [8]$$

where $\sigma(T)$ is the within-group standard deviation of true scores. Once this difference in defining ψ is taken into account, Equation 6 is mathematically equivalent to Equation 1. This can be seen by substituting

$$\rho = \sigma^2(T)/\sigma^2(X) \qquad [9]$$

for ϱ and by substituting Equation 8 for ψ into Equation 6, yielding Equation 1, where ψ is defined as in Equation 7. Thus, Subkoviak and Levin's formula is equivalent to the traditional formula; it is not the case that one is correct and the other incorrect. The formulas differ only because of the different definitions of ψ. In addition, it is possible to derive other formulas that are equivalent to these two. For example, ϕ might be defined as

$$\phi = \sqrt{((1-\rho)n\psi^2)/(\nu+1)\ (\Sigma a_k^2)} \qquad [10]$$

defining ψ as

$$\psi_E = (\Sigma a_k \mu_k)/\sigma(E) \qquad [11]$$

where $\sigma(E)$ is the within-group standard deviation of error scores.

Comparison of the Formulas

Which of these formulas is the appropriate one? According to Levin and Subkoviak (1977, p. 332), their formula (Equation 6 here) incorporating the reliability of the dependent measure should be used because Equation 1 results in "underestimates of required sample sizes (or overestimates of available power)" Forsyth (1978, p. 380), on the other hand, concluded that to use Equation 6, "data analyses would have to be performed on true scores rather than observed scores." Since analyses are actually done with observed scores, Forsyth concluded that Equation 1 should be used. However, both Subkoviak and Levin and Forsyth are mistaken because, as previously shown, the formulas are all equivalent. Thus, it is not the case that one formula is correct and the others incorrect, as these authors have improperly concluded. The only difference between the formulas arises from the manner in which ψ is defined; i.e., how one wishes to describe the magnitude of the treatment effect—in terms of the observed score standard deviation, or in terms of the true score standard deviation, or even in terms of the error score standard deviation.

Since the various formulas are all equivalent, it might seem that the choice of how to describe the magnitude of the treatment effect would not matter. However, this choice is in fact very important, both from a practical and a theoretical viewpoint. The reason is that the magnitude of group differences (ψ) is defined differently in each formula, and ψ must be specified prior to calculating ϕ. If ψ_X equals some constant d, then an equivalent ψ_T is necessarily larger than d for a fallible dependent measure. In fact, inspection of Equations 7 and 8 shows that ψ_T must equal $d/\sqrt{\varrho}$ if group separation is to be equivalent.

Consider, for example, three hypothetical researchers attempting to determine sample size, where it is known that the reliability of the dependent measure is 0.64. The first researcher decides to ignore reliability and hence uses Equation 1. He/she is interested in determining sample size for a magnitude of group separation given by $\psi_X = 1$, leading to a noncentrality parameter given by

$$\phi = \sqrt{n/(\nu+1)\ (\Sigma a_k^2)}. \qquad [12]$$

The second researcher decides to incorporate information concerning the reliability of the dependent measure and hence uses Equation 6. Defining ψ_T to be 1, seemingly comparable to the definition of the first researcher, results in

$$\phi = 0.8\sqrt{n/(\nu+1)\ (\Sigma a_k^2)}. \qquad [13]$$

The third researcher also incorporates the reliability information, but defines ψ_T to be 1.25, thus yielding

$$\phi = \sqrt{n/(\nu+1)\ (\Sigma a_k^2)}. \qquad [14]$$

For a fixed n, Researcher 1 seemingly has a more powerful statistical test than does Researcher 2. In fact, however, the real explanation of the different ϕ values is that the two ϕ values correspond to two different alternative hypotheses. Although the value of ψ in each case is 1, ψ_X here represents a larger group separation than does ψ_T, because σ_X is larger than σ_T. On the other hand, Researchers 1 and 3 have specified the same alternative hypothesis; hence the value of their noncentrality parameters is the same, despite the fact that their two ψ values are numerically different.

Thus, Equations 1, 6, and 10 are equivalent, but the attainment of numerically equivalent values of ϕ for the three formulas necessitates numerically different values of ψ, whenever $0<\varrho<1$. The real issue, then, is not which formula is correct but rather which definition of ψ is most appropriate, since stating that a meaningful alternative hypothesis for which power should be .80 is one corresponding to a ψ of 1 (for example) will in general yield different results for determining sample size, depending upon whether the ψ in question refers to ψ_X, ψ_T, or ψ_E. The reason that different results will be obtained is that a treatment effect of one standard deviation of observed scores is actually a larger effect than is an effect of one standard deviation of true scores and hence requires fewer subjects for a fixed level of power. This relationship can be seen by inspecting Equations 1 and 6, thus explaining Subkoviak and Levin's conclusion that Equation 1 underestimates the required number of subjects.

Alternate Definitions of ψ

Which definition of ψ seems to be most generally appropriate? If a researcher wants to detect a treatment effect of "w standard deviation units" (where w is some constant) or of "w raw score units," should these units refer to observed score units, true score units, or error score units? Although any of the three can be used to yield correct results, it is difficult to imagine a situation where the use of error score units would be meaningful. Hence, in practice, the choice seems to be between the traditional approach versus Subkoviak and Levin's approach.

Is it more meaningful to state magnitude of effect in terms of observed score units or in terms of true score units? Although the idea of using true score units sound appealing, it would seem that the use of observed score units is actually more meaningful for several reasons. First, researchers are much more accustomed to thinking of treatment effects in terms of observed scores, so that it is easier to specify a meaningful alternative hypothesis in terms of observed scores than in terms of true scores. Second, the inclusion of ϱ in the equation for ϕ introduces yet another parameter whose true population value must be estimated. Such errors in estimation will inevitably affect the accuracy of the actual sample size needed in an investigation. Third, and most important, it is possible that $\sigma(T)$ might

be near zero, or even theoretically be zero, yet group differences might be large relative to $\sigma(X)$ because $\sigma(T)$ refers to within-group individual differences, and measures sensitive to group differences may be insensitive to individual differences (see Nicewander & Price, 1978).

As an example, suppose that the dependent variable is a difference score whose components are highly correlated. In this situation it is well known that the reliability of the dependent variable is quite low. In fact, as the correlation between the components increases, the reliability of the difference score approaches zero. A researcher following Equation 6 (i.e., using Subkoviak and Levin's approach) will find that an enormous sample is required; indeed, as reliability approaches zero, the required sample size approaches infinity. However, Overall and Woodward (1977) have demonstrated that the power to detect group differences is actually increased as reliability of the difference score decreases, contrary to the conclusion given by Equation 6.

The apparent contradiction arises because what may seem to be a reasonable magnitude of treatment effect in terms of true score units reflects a miniscule effect in terms of observed score units, since $\sigma(T)$ is much smaller than $\sigma(X)$. This relationship will occur any time reliable individual differences within groups are small; in this situation, Subkoviak and Levin's approach is unduly pessimistic.

There seems to be no reason that the magnitude of between-group effects should be expressed in terms of reliable within-group individual differences variance, since the latter can conceivably be zero, yet treatment effects can exist. The problem seems to be that some researchers mistakenly believe that such a situation of no true variance within groups necessarily implies that any between-group differences are due to error and hence are not real. However, this is incorrect, as the example of difference scores shows, so that it is sensible to express magnitude of treatment effect in terms of the observed within-groups standard deviation. Traditional textbook formulas are not only appropriate but are also generally to be preferred to Subkoviak and Levin's formulation.

The Effect of Test Length on Power

Since traditional textbook formulas are appropriate and do not take reliability of the dependent measure into account, it might seem that reliability has no effect whatsoever upon power. In particular, it might seem that on the basis of the preceding discussion, lengthening a psychological test being used as the dependent variable in a study should have no effect on the power of the statistical test, although the reliability of the psychological test would be increased. Under certain reasonable assumptions, however, the power of the statistical test is increased by increasing test length.

Consider a set of I parallel items with a common intercorrelation of ρ and a common within-group variance of σ_i^2. It will further be assumed that the magnitude of the treatment effect is a constant for each item, i.e., there is a constant C such that for every item i

$$\sum_k a_k \mu_{ki} = C. \qquad [15]$$

Thus, for the test composed of I items,

$$\sum_i \sum_k a_k \mu_{ki} = IC, \qquad [16]$$

i.e., the magnitude of the effect for the total test score is I times C. The within-group variance of test scores is given by

$$I\,\sigma_I^2\,(1 + (I-1)\,\rho), \qquad [17]$$

since test scores are simply sums of scores on individual items. Substituting these expressions into Equation 1 for ϕ yields

$$\phi = \sqrt{nI^2C^2/I\sigma_I^2(1 + (I-1)\rho)\ (\nu+1)\ (\Sigma a_k^2)} \qquad [18]$$

since for the test as a whole,

$$\psi = IC/\sqrt{I\sigma_I^2\ (1 + (I-1)\ \rho)}\ . \qquad [19]$$

Simplification of Equation 18 leads to

$$\phi = \sqrt{nIC^2/\sigma_I^2(1 + (I-1)\rho)\ (\nu+1)\ (\Sigma a_k^2)}. \qquad [20]$$

Equation 20 is interesting for several reasons. First, it provides yet another approach for determining power. Once again, the distinction between this equation and the others for calculating power arises because of the definition of ψ. With Equation 20, ψ is expressed in terms of the magnitude of treatment effect for the observed score standard deviation of a single item. The sample size needed to attain a certain level of power for this value of ψ can be calculated if I and ϱ are specified. Second, unless $\varrho = 1$ (which corresponds to an infallible measure), ϕ increases as I increases, so that lengthening the dependent measure results in increased statistical power.

Third, the role of ϱ in affecting the extent to which lengthening the test increases power is made explicit. For example, if $\varrho = 1$, Equation 20 simplifies to

$$\phi = \sqrt{nC^2/\sigma_I^2\ (\nu+1)\ (\Sigma a_k^2)}, \qquad [21]$$

so that increasing the length of the test has no effect on power. Such a result is reasonable because ϱ will equal 1 if and only if the dependent measure is infallible. Thus, for an infallible variable, length of test is irrelevant. At the other extreme, if $\varrho = 0$, the equation simplifies to

$$\phi = \sqrt{nIC^2/\sigma_I^2\ (\nu+1)\ (\Sigma a_k^2)}, \qquad [22]$$

so that in terms of ϕ there is a direct trade-off between number of items and number of subjects.

Fourth, the formula makes explicit the relative effects on power of lengthening the test or increasing the sample size. Using Equation 20 it is possible to estimate the gain in power that would result from lengthening the test by some number of items and/or increasing the sample size by some number of subjects. In practice, the formula should be viewed as providing an estimate, since the assumption of parallel items may be violated, and since the researcher's judgments of C, ϱ, and σ_I^2 may be subject to error.

Nonetheless, inspection of Equation 20 reveals that increasing the sample size by some factor always results in a larger noncentrality parameter than would result from lengthening the test by that same factor, as long as ϱ is nonzero. If ϱ is zero, Equation 22 shows that the increase in ϕ is the same for an increase in sample size as for an equivalent increase in test length. However, number of subjects is actually more influential than number of items even here, since the degrees of freedom for the

denominator of the F test depends upon sample size but not upon number of items, and larger denominator df is associated with higher power.

These results agree with the findings of Overall and Dalal (1965) and Cleary and Linn (1969) that increasing sample size by some factor has a greater effect on power than does increasing test length by that same factor, regardless of the reliability of the test. Thus, if the cost of an additional measurement is constant, whether it is obtained by gaining an additional person or an additional item, it is better to increase sample size than test length. However, if additional subjects are more expensive than additional items, it may be more efficient to increase test length rather than sample size. (For further discussion, see Cleary and Linn, 1969, who developed equations that provide optimal sample size and test length subject to cost constraints.)

At this point, that length of test influences power with a fallible dependent measure may seem to contradict the fact that traditional formulas that ignore reliability are appropriate for determining power. Once again, the apparent paradox is a consequence of different definitions of ψ. If magnitude of effect is expressed in terms of the observed standard deviation of a single item, increasing the number of items results in increased power. However, if magnitude of effect is expressed in terms of observed standard deviation of the entire test, lengthening the test has no effect on power because the standard deviation of observed scores increases as the test becomes longer. In this sense, a standard deviation difference of 1 between groups on a 20-item test means that the groups are more different from one another than would a standard deviation difference of 1 on a 40-item test. Specifically, the magnitude of the effect for an individual item would be larger for the shorter test than for the longer test.

Hence, a researcher who uses a short test and specifies a ψ to be used in Equation 1 is actually expressing a larger treatment effect per item than is another researcher who specifies the same value for ψ but uses a longer test. Although Equation 1 would tell the two researchers that for a fixed n their power is the same, the researcher with the shorter test must actually have a larger treatment effect per item to achieve the same level of power as the researcher using the longer test. Again, Equations 1 and 20 are mathematically equivalent, differing only in the way in which ψ is defined.

Conclusion

As has been demonstrated, there are numerous correct methods for determining power. The choice between these methods depends upon how one wishes to express the magnitude of treatment effects. In other words, it is possible to state an alternative hypothesis against which power is to be judged in several different ways, and power varies with these different alternative hypotheses. Although all of the approaches are mathematically correct, the traditional approach seems most meaningful, and it is recommended that researchers continue to use this formula to determine power and required sample size. Nevertheless, researchers should be aware that there are definite benefits with respect to power of employing dependent measures that contain small error variance, such as can be obtained by lengthening the test on which measurements are taken.

References

Brewer, J. K. On the power of statistical tests in the *American Educational Research Journal*. *American Educational Research Journal*, 1972, *9*, 391-401.

Chase, L. J., & Chase, R. B. A statistical power analysis of applied psychological research. *Journal of Applied Psychology*, 1976, *61*, 234-237.

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

Cleary, T. A., Linn, R. L., & Walster, G. W. Effect of reliability and validity on power of statistical tests. In E. F. Borgatta & G. W. Bohrnstedt (Eds.), *Sociological methodology.* San Francisco: Jossey-Bass, 1970.

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

Forsyth, R. A. A note on "Planning an experiment in the company of measurement error" by Levin and Subkoviak. *Applied Psychological Measurement,* 1978, *2,* 379-383.

Glass, G. V., & Stanley, J. C. *Statistical methods in education and psychology.* Englewood Cliffs, NJ: Prentice-Hall, 1970.

Hays, W. L. *Statistics for the social sciences.* New York: Holt, Rinehart, & Winston, 1973.

Kirk, R. E. *Experimental design: Procedures for the behavioral sciences.* Belmont, CA: Brooks/Cole, 1968.

Levin, J. R. Determining sample size for planned and post hoc analysis of variance comparisons. *Journal of Educational Measurement,* 1975, *12,* 99-108.

Levin, J. R., & Subkoviak, M. J. Planning an experiment in the company of measurement error. *Applied Psychological Measurement,* 1977, *1,* 331-338.

Levin, J. R., & Subkoviak, M. J. Correcting "Planning an experiment in the company of measurement error." *Applied Psychological Measurement,* 1978, *2,* 384-387.

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

Overall, J. E., & Dalal, S. N. Design of experiments to maximize power relative to cost. *Psychological Bulletin,* 1965, *64,* 339-350.

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

Schmidt, F. L., Hunter, J. E., & Urry, V. W. Statistical power in criterion-related validation studies. *Journal of Applied Psychology,* 1976, *61,* 473-485.

Subkoviak, M. J., & Levin, J. R. Fallibility of measurement and the power of a statistical test. *Journal of Educational Measurement,* 1977, *14,* 47-52.

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

Winer, B. J. *Statistical principles in experimental design.* New York: McGraw-Hill, 1971.

Author's Address

Send requests for reprints or further information to Scott E. Maxwell, Department of Psychology, University of Houston, Houston, TX 77004.