Planning an Experiment in the Company of Measurement Error

Joel R. Levin and Michael J. Subkoviak
University of Wisconsin

"Textbook" calculations of statistical power and/or sample size follow from formulas that assume that the variables under consideration are measured without error. However, in the "real world" of behavioral research, errors of measurement cannot be neglected. A recent sample-size determination approach is easily adapted to incorporate unreliability information for both completely randomized and randomized block analysis-of-variance designs. A worked example presents an instance wherein a blocking strategy is clearly advantageous assuming infallible measuring instruments, but not when the same instruments are granted fallibility.

When it comes to designing an experiment, a behavioral researcher can draw from a variety of sources—some in the form of old wives' tales and some in the form of theoretically sound recommendations (e.g., Feldt, 1958)—to determine whether it is preferable to assign subjects randomly to K experimental conditions and subsequently to perform an analysis of variance on the dependent variable Y (hereafter referred to as a completely randomized design) or whether to include in the analysis antecedent information based on variable X (known or assumed to be related to Y). The antecedent information included can be operationally dealt with in various ways: in terms of randomized blocks analysis, analysis of covariance, or analysis of an index of response (such as change scores)—cf. Porter & Chibucos (1974).

The major advantage of these procedures, relative to the completely randomized design, is one of reducing the within-treatment variability by removing the variation in Y that is due to the relationship between X and Y. This paper focuses on one of these procedures, namely the randomized block design, as a competitor to the completely randomized design; and, in particular, it considers an alternative to the traditional way of deciding whether or not to block and includes real-life situations in which errors of measurement associated with X, Y, or both are likely to be present. Moreover, since the discussion by Porter and Chibucos (1974) suggests that in "true" (Campbell & Stanley, 1966) experiments of moderate sample size, analysis of covariance and analysis of an index of response may be regarded as essentially equivalent procedures to blocking. Allowing for degrees-of-freedom differences and slight differences in their error expected mean squares, the material presented here has implications for the other two procedures as well.[1]

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 1, No. 3 Summer 1977 pp. 331–338
© Copyright 1977 West Publishing Co.

[1] The present discussion focuses on the classical randomized block design in which only one subject within each block is assigned to each treatment condition (i.e., n = 1). It is with

Reliability and Sample Size

Statistics texts typically acknowledge four ingredients of hypothesis testing: *a)* Type I error probability (α); *b)* Type II error probability (β) or its complement, power (1 - β); *c)* sample size; and *d)* the magnitude of the experimental effect of interest. In planning an experiment, a researcher can specify α and the power desired to detect an effect of specified magnitude and subsequently calculate the required sample size (or, in evaluating an experiment, the predetermined α level and sample size can be used to compute the power available to detect an effect of given magnitude).

Such calculations tacitly assume that dependent variables and/or antecedent variables are measured without error, i.e., they are perfectly reliable *(true scores).* In actual practice, however, both antecedent and dependent variables are likely to be measured with error, i.e., they are fallible *(observed scores).* The result is that "textbook" power/sample size calculations which do not take the unreliability of the observed data into account will produce inaccurate estimates. In particular, they will produce underestimates of required sample sizes (or overestimates of available power). This paper demonstrates the necessity for behavioral researchers to pay explicit attention to the issue of measurement error when planning an experiment. To assist in this endeavor, procedures for computing power and/or sample size are provided that include the reliability coefficient of observed scores as a fifth hypothesis-testing ingredient.

It is possible, of course, to obtain variables with little or no measurement error associated with them. For example, one could include "weight" as both an antecedent and a dependent variable to study the effect of various diets or one could block on "family size" to determine the consequences of various intervention programs. Here, however, arguments will be developed for variables possessing less than perfect reliability, with perfectly reliable antecedent and/or dependent variables falling out as special cases.

Several authors have considered the effect of unreliability on statistical tests (e.g., Cleary & Linn, 1969; Cleary, Linn, & Walster, 1970; Overall & Dalal, 1965; Porter, 1967; Sutcliffe, 1958). Cleary *et al.* (1970), for example, have demonstrated that the power of the F-test in a one-way, fixed-effects analysis of variance (ANOVA) decreases as the reliability—and also as the validity—of the dependent variable decreases. One purpose of the present paper is to extend some of the Cleary *et al.* notions to designs in which antecedent information is considered (in particular, to the randomized block design). Moreover, in contrast to the commonly recommended strategy for deciding whether or not it would be advantageous to block [i.e., by determining the relative efficiency of a randomized block design to a completely randomized design for a fixed number of subjects (cf. Kirk, 1968, pp. 147–149)], the strategy adopted here consists of framing the decision in terms of the respective sample sizes associated with the two designs that are required to yield equivalent power for detecting specified effects of interest (see, for example, Cohen, 1969, pp. 46–50).

Case 1: Latent True Variables

Sample Size Determination for the Completely Randomized Design

Levin (1975) discusses sample size determination based on a researcher's *a priori* specification of the minimum value of any given linear contrast of interest (which has been called $\dot{\Psi}_o$) in accordance with desired α and 1-β. The resulting number of subjects required per experimental condition (assuming $N_1 = N_2 = \ldots = N_K = n$) guarantees the researcher the desired power to

this design and its assumption of within-block homogeneity on the antecedent variable that the above "equivalence" statement holds. For the generalized randomized block design with more than one subject in each block-treatment combination (i.e., n > 1) and/or when within-block heterogeneity is present, other considerations such as the number of treatments, blocks, and subjects included become relevant (cf. Feldt, 1958).

detect a contrast as large or larger in magnitude than that specified. In the case of a planned-comparison approach to hypothesis testing, an F-test of the contrast is performed with 1 and $K(n\text{-}1)$ degrees of freedom (these referring to the degrees of freedom associated with the contrast and the mean square within respectively); and in this situation the probability of detecting a contrast of the magnitude specified is alternatively the probability of obtaining a significant F-ratio (both 1 - β). In the case of a *post hoc* approach to hypothesis testing, an omnibus F-test is performed with K-1 and $K(n\text{-}1)$ degrees of freedom (where K - 1 represents the degrees of freedom associated with the mean square between ν_1); and in this situation the probability of detecting a contrast of the magnitude specified is alternatively the probability of obtaining a significant F-ratio and then identifying that contrast as statistically significant according to Scheffé's (1953) multiple comparison procedure (see Levin, 1975). According to this formulation, Ψ_σ represents the magnitude of the contrast in means considered to be of interest to the researcher and which is expressed in within-treatment standard deviation units (σ). Thus if

$$\Psi = \sum_{k=1}^{K} a_k \mu_k \qquad [1]$$

(where the a_k represent contrast coefficients chosen such that $\sum_{k=1}^{K} a_k = 0$), then

$$\Psi_\sigma = \frac{\sum_{k=1}^{K} a_k \mu_k}{\sigma} \, . \qquad [2]$$

Sample Size Determination for the Randomized Block Design

Rather than adopting the completely randomized design, a researcher may choose to form n blocks of K subjects (on the basis of some relevant antecedent information) and then randomly assign subjects within blocks to the K treatment conditions. It is well known that the effect of introducing a blocking variable into the design reduces σ by a factor of $\sqrt{1 - \varrho^2_{XY}}$, where ϱ_{XY} represents the correlation between the antecedent variable and the dependent variable (see, for example, Feldt, 1958). Thus, in terms of the present approach, all that needs to be done is to redefine a standardized contrast as

$$\Psi^*_\sigma = \frac{\sum_{k=1}^{K} a_k \mu_k}{\sigma\sqrt{1 - \rho^2_{XY}}} = \frac{\Psi_\sigma}{\sqrt{1 - \rho^2_{XY}}} \, . \qquad [3]$$

The effect of blocking, then, is to increase the value of Ψ_σ of the completely randomized design. If this increase overcompensates for the corresponding loss in error degrees of freedom, i.e., from $K(n$ - $1)$ to $(K$ - $1)$ $(n$ - $1)$, then there will be a decrease in the number of subjects required to maintain equivalent power to that in the completely randomized case.

Case 2: Fallible Variables

The above discussion has proceeded under the assumption that the only "error" in the ANOVA model consists of subject error. If there is measurement error as well, one's effective power will not be so great as one's nominal power; or, stated differently, a researcher will require more subjects than the "textbook" sample size determination indicates are needed in order to have the desired power (see, for example, Cleary *et al.*, 1970). Classical test theory (Lord & Novick, 1968) assumes that the observed score Y_i for person i is equal to his or her true score T_i plus measurement error E_i, such that $Y_i = T_i + E_i$. Since T_i and E_i are assumed to be independently distributed with respective expected values of μ_T and 0 and respective variances of σ^2_T and σ^2_E, it follows that:

$$\begin{aligned} \mu_Y &= \mu_T + 0 \\ &= \mu_T \end{aligned} \qquad [4]$$

and

$$\sigma^2_Y = \sigma^2_T + \sigma^2_E \, . \qquad [5]$$

The reliability of observed scores Y_i is the ratio of true score variance to observed score variance:

$$\rho_{YY'} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2} = \frac{\sigma_T^2}{\sigma_Y^2} . \qquad [6]$$

Sample Size Determination for the Completely Randomized Design

How do these properties affect sample size determination in the completely randomized design? As was noted previously, Ψ_σ is simply a contrast involving the treatment means which is expressed in within-treatment standard deviation units. Because of the relationship in Equation 4, the numerator of Ψ_σ is unaffected by measurement errors. What *is* affected is the denominator. Thus, σ in Ψ_σ reflects the within-treatment standard deviation of true scores, or σ_T. Following Cleary *et al.* (1970) and employing Equation 6, we note that in terms of observed scores,

$$\sigma_Y = \frac{\sigma_T}{\sqrt{\rho_{YY'}}} . \qquad [7]$$

Thus, for the usual case where measurement errors associated with the dependent variable are expected, we simply redefine Ψ_σ as:

$$\underline{\Psi}_\sigma = \frac{\sum_{k=1}^{K} a_k \mu_k}{\sigma_T / \sqrt{\rho_{YY'}}} = \sqrt{\rho_{YY'}}\ \Psi_\sigma \qquad [8]$$

where it may be easily shown that $\varrho_{YY'}$ represents the (assumed common) *within-treatment* reliability of the dependent variable (see Subkoviak & Levin, in press).

In the case of the randomized block design, the numerator of Ψ_σ^* is again unaffected by measurement errors in X and Y, while the quantities σ and $\sqrt{1 - \varrho_{XY}^2}$ in the denominator are both affected. Thus, if T_X and T_Y are used to symbolize the true score parts of X and Y, Ψ_σ^* can be redefined (in the presence of measurement error) as:

$$\underline{\Psi}_\sigma^* = \frac{\sum_{k=1}^{K} a_k \mu_k}{\sigma_Y \sqrt{1 - \rho_{XY}^2}}$$

$$= \left[\frac{\sigma_{T_Y}}{\sigma_{T_Y}} \cdot \frac{\sqrt{1 - \rho_{T_X T_Y}^2}}{\sqrt{1 - \rho_{T_X T_Y}^2}} \right] \frac{\sum_{k=1}^{K} a_k \mu_k}{\sigma_Y \sqrt{1 - \rho_{XY}^2}}$$

$$= \left[\frac{\sigma_{T_Y}}{\sigma_Y} \cdot \frac{\sqrt{1 - \rho_{T_X T_Y}^2}}{\sqrt{1 - \rho_{XY}^2}} \right] \frac{\sum_{k=1}^{K} a_k \mu_k}{\sigma_{T_Y} \sqrt{1 - \rho_{T_X T_Y}^2}}$$

$$= \left[\sqrt{\rho_{YY'}} \cdot \frac{\sqrt{1 - \rho_{XY}^2 / (\rho_{XX'} \rho_{YY'})}}{\sqrt{1 - \rho_{XY}^2}} \right] \Psi_\sigma^*$$

$$= \sqrt{\frac{\rho_{XX'} \rho_{YY'} - \rho_{XY}^2}{\rho_{XX'} (1 - \rho_{XY}^2)}}\ \Psi_\sigma^* \qquad [9]$$

where $\varrho_{XX'}$ and $\varrho_{YY'}$ are the respective reliabilities of X and Y, and $\varrho_{T_X T_Y}^2 = \varrho_{XY}^2 / (\varrho_{XX'} \varrho_{YY'})$ follows from the correction-for-attenuation formula. (As will become apparent in the example presented in the next section, in practice one need not "know" the value of $\varrho_{T_X T_Y}$. Rather, Equation 8 may be used in conjunction with reasonable estimates of the observed within-treatment quantities, ϱ_{XY}, $\varrho_{XX'}$, and $\varrho_{YY'}$.)

It should be noted that this expression can be easily adapted to fit various special cases. In particular, if only X is assumed to be fallible, it may be seen that:

$$\underline{\Psi}_\sigma^* = \sqrt{\frac{\rho_{XX'} - \rho_{XY}^2}{\rho_{XX'} (1 - \rho_{XY}^2)}}\ \Psi_\sigma^* . \qquad [10]$$

On the other hand, if only Y is fallible:

$$\underline{\Psi}^*_\sigma = \sqrt{\frac{\rho_{YY'} - \rho^2_{XY}}{1 - \rho^2_{XY}}}\ \Psi^*_\sigma \ . \qquad [11]$$

Finally, if neither X nor Y is fallible:

$$\underline{\Psi}^*_\sigma = \sqrt{\frac{1 - \rho^2_{XY}}{1 - \rho^2_{XY}}}\ \Psi^*_\sigma = \Psi^*_\sigma \qquad [12]$$

which is as it should be.

An Example

Levin's (1975) sample size determination formula is given by:

$$\phi = \sqrt{\frac{n\Psi^2_\sigma}{(\nu_1+1)\sum_{k=1}^{K} a^2_k}} \qquad [13]$$

where: ϕ = a parameter in the Pearson and Hartley (1951) power charts, available in most experimental design textbooks; more complete tables displaying ϕ are also available (e.g., Tiku, 1967, 1972).

Let us apply Equation 13 to the simplest ANOVA situation, namely for $K = 2$ which is equivalent to the independent two-sample nondirectional t-test situation.

Assume that a researcher wishes to have an 80 percent chance of detecting a difference in $K = 2$ means of at least 1 standard deviation unit, how many subjects per treatment group should he/she include, based on a Type I error probability of .05? [With reference to Equation 13, it should be noted that $\nu_1 + 1$ will always equal K in the one-way layout (here $\nu_1 + 1 = 2$); and when only pairwise differences in means are of interest, $\sum_{k=1}^{K} a^2_k = 2$. However, in some situations, complex comparisons may interest the researcher, in which case the value of $\sum_{k=1}^{K} a^2_k$ will change (see Levin, 1975).]

The information contained in the preceding paragraph may be translated as follows: $\alpha = .05$, $1 - \beta = .80$, $\Psi_\sigma = 1.00$. Incorporating this into Equation 13 and the appropriate power charts and proceeding in the manner described by Levin, we find that in the completely randomized situation (assuming a perfectly reliable dependent variable), a total of 17 subjects per treatment group is required to yield the desired power.

If we further assume that an antecedent variable is selected that correlates .50 with performance on the dependent measure (i.e., $\varrho_{XY} = .50$), then it can be seen that $\Psi^*_\sigma = 1/\sqrt{1-(.50)^2} = 1.155$. Substituting this into Equation 13 and checking with the appropriate ν_2, we find that if the randomized block design were employed (and assuming perfectly reliable antecedent and dependent variables), a total of 14 blocks ($n = 14$) would be required to yield equivalent power to that in the completely randomized design above.

Now let us suppose that either or both of the two variables involved (antecedent and dependent) are fallible. Given separate (and equal) reliabilities of $\varrho_{XX'} = \varrho_{YY'} = .80$, for example, we are able to retrace the steps associated with Equation 13, incorporating Ψ_σ and Ψ_σ as previously defined. Table 1 summarizes the results of this endeavor.

What is especially interesting about this example is that even though we start out with a situation in which it is clearly preferable to block (as reflected by a total savings of six subjects for Situation 1 of Table 1), the randomized block advantage disappears (as reflected by the 0 total subject savings difference in Situation 4 of Table 1) by the time the antecedent and dependent variables are both granted fallibility on the order of $\varrho_{XX'} = \varrho_{YY'} = .80$.

To make this lesson somewhat more concrete, assume that a researcher is interested in comparing the efficacy of two instructional varia-

Table 1

Comparison of Completely Randomized (CR) and Randomized Block (RB) Design Sample Sizes for the Present Example (K = 2, α = .05, 1 - β = .80, Ψ_σ = 1.00, ρ_{XY} = .50)

Situation	Ψ_σ or Equivalent	Number of Subjects Per Group	Total Subject Savings [2(CR-RB)]
1. X is Infallible, Y is Infallible			
CR	1.000	17	6
RB	1.155	14	
2. X is Infallible, Y is Fallible ($\rho_{YY'}$ = .80)			
CR	.894	21	6
RB	.989	18	
3. X is Fallible ($\rho_{XX'}$ = .80), Y is Infallible			
CR	1.000	17	4
RB	1.106	15	
4. X is Fallible ($\rho_{XX'}$ = .80), Y is Fallible ($\rho_{YY'}$ = .80)			
CR	.894	21	0
RB	.931	21	

tions designed to teach eighth-grade mathematics. Both variations are to be incorporated into programmed instruction booklets and randomly assigned to students within classrooms or schools), and end-of-year performance will be assessed via a standardized mathematics achievement test. Suppose further in this hypothetical situation that the production cost of the booklets is somewhat of a factor; in this case, an experimental design that will yield the desired power with the fewest students is the one to be selected. Given this information, should the researcher randomly assign students to the two treatment conditions or block on seventh-grade standardized mathematics achievement scores found empirically (based on a review of the literature and/or pilot research) to be correlated .50 with eighth-grade scores? Ignoring the unreliability associated with two achievement tests (as in the "textbook" case), the researcher would clearly do well to block; he/she would require six fewer students with a randomized block design than with a completely randomized design. However, considering that the published reliabilities of the two tests were .80, the researcher would discover that it makes little difference which of the two experimental designs he/she selects, since there is a 0 subject savings. In fact, if it would require some additional effort to obtain and/or record the seventh-grade achievement data, the researcher may well opt for the seemingly less efficient (though not so in this case), completely randomized design.

Conclusion

This example is but one of several that could have been contrived to demonstrate the following points. First, each potential experiment should be examined on an *a priori* basis to determine whether or not it is advantageous to block. This decision cannot be made without considering the number of treatment conditions included, the magnitude of the relationship between the antecedent and blocking variables (ϱ_{XY}), as well as the various hypothesis-testing ingredients described at the outset of the paper. Second, to follow these procedures without simultaneously considering errors of measurement is to live in a "fool's paradise," for these, too, will affect block/no-block decisions. In cases where *a priori* reliability information is lacking, pilot research or sagacious judgements (to obtain approximate and conservative estimates, respectively) will surely do better than nothing.

References

Campbell, D. T., & Stanley, J. C. *Experimental and quasi-experimental designs for research.* Chicago: Rand, McNally, 1966.

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 and G. W. Bohrnstedt (Eds.), *Sociological methodology.* San Francisco: Jossey-Bass, 1970.

Cohen, J. *Statistical power analysis for the behavioral sciences.* New York: Academic Press, 1969.

Feldt, L. S. A comparison of the precision of three experimental designs employing a concomitant variable. *Psychometrika,* 1958, *23,* 335–353.

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.

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

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

Pearson, E. S., & Hartley, H. O. Charts of the power function for analysis of variance tests, derived from the non-central *F*-distribution. *Biometrika,* 1951, *38,* 112–130.

Porter, A. C. *The effects of using fallible variables in the analysis of covariance.* Unpublished doctoral dissertation, University of Wisconsin, Madison, 1967.

Porter, A. C., & Chibucos, T. R. Analysis issues in summative evaluation. In G. Borich (Ed.), *Evaluating educational programs and products.* Englewood Cliffs, NJ: Educational Technology Press, 1974.

Scheffé, H. A method for judging all contrasts in the analysis of variance. *Biometrika*, 1953, *40*, 87–104.

Subkoviak, M. J., & Levin, J. R. Fallibility of measurement and the power of a statistical test. *Journal of Educational Measurement*, in press.

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

Tiku, M. L. Tables of the power of the *F*-test. *Journal of the American Statistical Association*, 1967, *62*, 525–539.

Tiku, M. L. More tables of the power of the *F*-test. *Journal of the American Statistical Association*, 1972, *67*, 709–710.

Author's Address

Joel R. Levin, Department of Educational Psychology, University of Wisconsin, 1025 West Johnson Street, Madison, Wisconsin 53706.