A Note on "Planning an Experiment in the Company of Measurement Error" by Levin and Subkoviak

Robert A. Forsyth
University of Iowa

Levin and Subkoviak (1977) contend that experimenters should consider the reliability of their measures when comparing the power of fixed-effects completely randomized designs with that of randomized block designs. The two major purposes of this note are (1) to show that under the conditions specified by Levin and Subkoviak, it is not necessary to consider the reliabilities when comparing these two designs and (2) to identify certain errors in the illustrative example used by Levin and Subkoviak.

Levin and Subkoviak (1977) recently considered the relative power of the fixed-effects completely randomized design (CRD) as compared with the randomized block design (RBD) when errors of measurement are present in the blocking (antecedent) variable and/or the dependent measure. They contended that the reliabilities of these measures must be considered when comparing the power of these two designs. Using a procedure for estimating power which was outlined by Levin (1975) and employing a specific numerical example, they purported to show that although the RBD was superior (i.e., fewer subjects were needed for a given degree of power) when the two variables were assumed infallible, the "randomized block advantage disappears . . . by the time the antecedent and dependent variables are both granted fallibility on the order of $\varrho_{xx'}$ [reliability of blocking variable] = $\varrho_{yy'}$ [reliability of dependent measure] = .80" (p. 335).

The two major purposes of this note are (1) to show that under the conditions specified in the Levin and Subkoviak article (hereafter referred to as L&S), it is not necessary to consider the reliabilities of the measures when comparing the power of the CRD with that of the RBD[1] and (2) to identify certain errors in the illustrative example used by L&S. When these errors are corrected, the results of the illustrative example do not support the conclusion that the advantage of the RBD disappears when the reliabilities of the antecedent and dependent measures are both assumed to be .80.

When the Levin (1975) procedure for determining the sample size for a CRD is used, the researcher must specify Ψ_o, which "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)" (p. 333). For the CRD, we have

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 2, No. 3 Summer 1978 pp. 377–381

[1]It should be noted that the L&S article is concerned with "the classical randomized block design in which only one subject within each block is assigned to each treatment condition" (p. 331).

$$\Psi_{\sigma} = \frac{\sum_{k=1}^{K} a_k \mu_k}{\sigma}$$

(L & S Equation 2)

where K = number of groups;
μ_k = mean of population k; and
$\sum_{k=1}^{K} a_k \mu_k$ = the contrast of interest to the researcher
$\sum_{k=1}^{K} a_k = 0$.

For the RBD, Ψ_{σ}^{*} is defined as follows:

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

(L & S Equation 3)

where ρ_{XY} represents the correlation between the antecedent variable and the dependent variable. Later in their article, L&S note that the σ in Ψ_{σ} "reflects the within-treatment standard deviation of true scores, or σ_T" (p. 334)[2].

[2]This decision to use σ_T to define Ψ_{σ} and Ψ_{σ}^{*} is completely arbitrary. It would have been just as "reasonable" to define Ψ_{σ} and Ψ_{σ}^{*} in terms of the common within-treatment observed score variability, σ_Y. In fact, most discussions related to the power/sample size issue involve quantities defined in terms of σ_Y (or σ_Y^2) rather than σ_T (or σ_T^2). (See, for example, Cohen, 1977; Levin, 1975; Myers, 1972; Kirk, 1968.) L&S (p. 332) contend that these power/sample size discussions "tacitly assume that the dependent and/or antecedent variables are measured without error . . . ." This contention does not seem to be justified. The use of σ_Y^2 rather than σ_T^2 in power/sample size determinations does not preclude the existence of mesurement error. Since, under the usual classical test theory assumption, the observed score variance (σ_Y^2) is assumed to be equal to the true score variance (σ_T^2) plus the measurement error variance (σ_E^2), the use of σ_Y^2 in these power/sample size calculations very definitely allows for the possibility of measurement error (i.e., that $\sigma_E^2 > 0$).

Thus, rewriting the above two equations to reflect this scaling in terms of σ_T, we have

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

$$\Psi_{\sigma}^{*} = \frac{\sum_{k=1}^{K} a_k \mu_k}{\sigma_T\sqrt{1 - \rho_{T_X T_Y}^2}} \qquad [2]$$

where $\rho_{T_X T_Y}$ represents the correlation between true scores on variables X and Y. The Ψ_{σ} and Ψ_{σ}^{*} values can then be used in what L&S call the sample size determination formula. This formula is shown below for Ψ_{σ}:

$$\phi = \sqrt{\frac{n\Psi_{\sigma}^2}{(\nu_1 + 1)\sum_{k=1}^{K} a_k^2}}$$

(L & S Equation 13)

where ϕ = a parameter in the Pearson and Hartley (1951) power charts and
ν_1 = the numerator degrees of freedom associated with the F-test to be performed.

Levin (1975) illustrates the use of this formula.

Using the above equation with the Ψ_{σ} and Ψ_{σ}^{*} values, the sample size necessary for a given power can be determined for each design. Of course, since the Ψ_{σ} and Ψ_{σ}^{*}values are defined in terms of true score variance, the subsequent data analyses would have to be performed on true scores rather than observed scores to achieve the designated power for a specified contrast, $\Sigma a_k \mu_k$, with each design. Since it is highly unlikely that any researcher would be using "true scores" in an analysis, using the Ψ_{σ} and Ψ_{σ}^{*} values defined by Equations 1 and 2 above in the

sample size determination formula provides little help to the experimenter. To estimate the sample size necessary for a given degree of power when observed scores are analyzed, it is necessary to define the contrasts of interest in terms of σ_Y rather than σ_T. If the contrasts are scaled in terms of σ_Y, we have

$$\underline{\Psi}_{\sigma} = \frac{\Sigma a_k \mu_k}{\sigma_Y} \qquad [3]$$

$$\underline{\Psi}^*_{\sigma} = \frac{\Sigma a_k \mu_k}{\sigma_Y \sqrt{1 - \rho^2_{XY}}} \qquad [4]$$

As before, the numerical values for $\underline{\Psi}_{\sigma}$ and $\underline{\Psi}^*_{\sigma}$ can be used in the sample size determination formula shown above. Now, however, the derived sample size represents the sample size for a given power and a specific $\Sigma a_k \mu_k$ when the data analysis is performed on the observed scores.

It is really the $\underline{\Psi}_{\sigma}$ and $\underline{\Psi}^*_{\sigma}$ values that L&S use to compare the RBD with the CRD for a specific set of parameters. However, they first derive definitions of $\underline{\Psi}_{\sigma}$ and $\underline{\Psi}^*_{\sigma}$ that are algebraically equivalent to Equations 3 and 4, but which explicitly involve the reliabilities of the X and Y measures in the formulas. These alternative expressions are based on the usual classical test theory assumptions that an examinee's observed score (Y) is equal to his or her true score plus measurement error, that is, $Y_i = T_i + E_i$. Given the usual assumptions, it also follows that $\sigma^2_Y = \sigma^2_T + \sigma^2_E$. These alternative expressions of $\underline{\Psi}_{\sigma}$ and $\underline{\Psi}^*_{\sigma}$ are also based on the classical definition of reliability: $\rho_{YY'} = \sigma^2_T/\sigma^2_Y$. Or, $\sigma^2_Y = \sigma^2_T/\rho_{YY'}$. Given the above, we have

$$\underline{\Psi}_{\sigma} = \frac{\Sigma a_k \mu_k}{\sigma_Y} = \frac{\Sigma a_k \mu_k}{\sigma_T/\sqrt{\rho_{YY'}}} = \sqrt{\rho_{YY'}}\ \Psi_{\sigma}$$

(L & S Equation 8)

Likewise, it can be shown (see L&S, p. 334) that

$$\underline{\Psi}^*_{\sigma} = \sqrt{\frac{\rho_{XX'}\rho_{YY'} - \rho^2_{XY}}{\rho_{XX'}(1 - \rho^2_{XY})}}\ \Psi^*_{\sigma}$$

(L & S Equation 9)

These last two equations are used by L&S to compute $\underline{\Psi}_{\sigma}$ and $\underline{\Psi}^*_{\sigma}$ in order to compare the RBD with the CRD. Note that L&S Equations 8 and 9 are algebraically equivalent to Equations 3 and 4, respectively. Thus, for fixed values of ϱ_{XY} and Ψ_{σ} (or $\underline{\Psi}_{\sigma}$), the relative merits of the RBD as compared to the CRD should be the same, regardless of which pair of equations (L&S 8 and 9 or Equations 3 and 4) are used to define $\underline{\Psi}_{\sigma}$ and $\underline{\Psi}^*_{\sigma}$. Since Equations 3 and 4 do not explicitly involve the reliabilities of X and Y, it is not clear why the reliabilities should directly affect the comparisons of the two designs. The L&S illustrative example, which fixes ϱ_{XY} and Ψ_{σ}, seems to indicate that the reliabilities do affect these comparisons; but, as noted previously, the illustrative example used by L&S contains some errors which account for the seemingly inconsistent results they report.

The numerical example in L&S compares the sample sizes needed for each design (CR and RB), given the following conditions:

$K = 2$
$\alpha = .05$
$\beta = .20$
$\varrho_{XY} = .50$

L&S consider the four situations outlined in the first column of Table 1. The title and column headings of Table 1 are identical with the title and column headings of Table 1 in L&S.

Before proceeding, it should be noted that the heading of column 2 is not accurate. The values in column 2 are expressed in terms of observed score variance rather than true score variance. (Of course, this is not a problem in Situation 1.) Therefore, these values should be labeled "$\underline{\Psi}_{\sigma}$ or Equivalent" rather than "Ψ_{σ} or Equivalent."

Table 1

Comparison of Completely Randomized (CR) and Randomized Block (RB) Design Sample Sizes for the Present Example ($K = 2$, $\alpha = .05$, $1 - \beta = .80$, $\Psi_\sigma = 1.00$, $\rho_{XY} = .50$)

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 (1.000)*	17 (17)*	6 (6)*
RB	1.155 (1.155)	14 (14)	
2. X is Infallible, Y is Fallible ($\rho_{YY'} = .80$)			
CR	.894 (.894)	21 (21)	8 (6)
RB	1.0328 (.989)	17 (18)	
3. X is Fallible ($\rho_{XX'} = .80$), Y is Infallible			
CR	1.000 (1.000)	17 (17)	6 (4)
RB	1.155 (1.106)	14 (15)	
4. X is Fallible ($\rho_{XX'} = .80$), Y is Fallible ($\rho_{YY'} = .80$)			
CR	.894 (.894)	21 (21)	8 (0)
RB	1.0328 (.931)	17 (21)	

*Values given in L & S.

The errors in the L&S example occurred in the calculation of $\underline{\Psi}^*_o$ for Situations 2, 3, and 4. The correct values for $\underline{\Psi}^*_o$ defined by L&S Equation 9 are shown in column 2. The values reported in the L&S article are shown in parentheses. It appears that the L&S values for $\underline{\Psi}_o$ in column 2 were obtained by defining Ψ^*_o incorrectly as equal to $\Psi_o/(\sqrt{1 - \rho^2_{xy}})$ rather than $\Psi_o/\sqrt{1 - \rho^2_{T_XT_Y}}$. However, in their derivation of Equation 9, it is clear that L&S define Ψ^*_o as equal to the latter quantity.

The most obvious conclusion from the recomputed values in column 2 is that the RBD advantage does not disappear when "the antecedent and dependent variables are both granted fallibility on the order of $\rho_{XX'} = \rho_{YY'} = .80$." In fact, on the basis of this one example, just the opposite conclusion seems justified. As can be seen in column 4, the RBD results in a savings of approximately eight subjects when both variables have a reliability of .80, as opposed to a savings of approximately six subjects when both variables are considered infallible. However, this comparison of Situations 1 and 4 is somewhat ambiguous, since the degree of falsity of H_o (in terms of $\underline{\Psi}_o$) deemed important is not comparable for the two situations. It is true, however, that the ratio of $\underline{\Psi}^*_o$ to $\underline{\Psi}_o$ will always be 1.155 under the conditions of this example, regardless of the reliabilities of X and Y. In general, for a given value of Ψ_o (or $\underline{\Psi}_o$) and a given value of ρ_{XY}, the ratio of $\underline{\Psi}^*_o$ to $\underline{\Psi}_o$ is always equal to $1/\sqrt{1 - \rho^2_{XY}}$, regardless of the reliabilities of the X and Y variables.

In summary, if the researcher identifies a contrast of interest (scaled either in terms of σ_Y or σ_T) and specifies a value for ρ_{XY}, then it is not necessary to consider the reliabilities of the X and Y measures when comparing the relative merits of the CRD and the RBD. In fact, introducing the reliability issue seems to make a fairly complicated problem unnecessarily even more complicated.

References

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

Kirk, R. E. *Experimental design procedures for the behavioral sciences.* Belmont, CA: Brooks/Cole Publishing Company, 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.

Myers, J. L. *Fundamentals of experimental design* (2nd ed.). Boston, MA: Allyn & Bacon, Inc., 1972.

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.

Author's Address

Robert A. Forsyth, 318 Lindquist Center for Measurement, University of Iowa, Iowa City, IA 52242