Correcting "Planning an Experiment in the Company of Measurement Error"

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

Comments on our earlier article are acknowledged and appreciated. In addition, potentially misleading notions arising from these comments are addressed and clarified.

Forsyth's (1978) comments serve to clarify several specifics of our 1977 article concerning the effect of measurement error on a researcher's choice of an analysis-of-variance design strategy. Among the most important is our hypothetical example in which an attenuated correlation (between the blocking variable, X, and the dependent variable, Y) should have been applied to the randomized block design for Situations 2 to 4 but was not. For this discovery by Forsyth we are grateful, since the recomputed values in his Table 1 are clearly different from ours. As will be shown, however, he appears to have overstated his case with respect to conclusions that may legitimately be drawn from that particular example.

At the same time, we find that we had overstated (even misstated) our case also, and for this we apologize. It is true that if an investigator chooses to think of standardized mean differences in terms of *observed* scores and their associated variability (σ_Y^2), he/she is certainly entitled to do so. (In such instances, the regular textbook version of our sample-size/power formula, Levin, 1975, will be appropriate.[1] In switching from a completely randomized design to a randomized block design, a researcher need only include the factor $\sqrt{1-\rho_{XY}^2}$ in the denominator of Ψ_o and proceed with appropriately modified degrees of freedom.)

There are those, however, who may prefer to approach the sample-size determination problem in terms of *true* scores and their associated variability ($\sigma_{T_Y}^2$). In this regard, we were somewhat surprised that Forsyth (1978) did not see fit to track down—or, at least, to cite—the prior literature on the same topic (e.g., Cleary & Linn, 1969; Cleary, Linn, & Walster, 1970; Sutcliffe, 1958). Had he sought out these references which were included in our article, he would have discovered that we started with the well-known relationship between test reliability and statistical power in the completely randomized design and simply generalized these notions to a blocking design.[2] Since in the past these authors and others have viewed the topic as a relevant one,

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 2, No. 3 Summer 1978 pp. 382–385

[1]We are grateful to Andrew C. Porter for some previous exchanges we have had on this issue.

[2]In defense of our Equations 8 and 9, for which Forsyth provides algebraic equivalents, it should be noted that these were intentionally derived to be consistent with previous formulas which include explicit reliability information (e.g., Cleary et al., 1970).

our starting assumption of a researcher wishing to consider true-score differences would seem to command at least some modest amount of attention from behavioral researchers, statisticians, and psychometricians.

Should a researcher even choose to *think* in terms of observed-score differences, however, Forsyth's (1978) concluding remarks are apt to leave the mistaken impression that reliability is a matter of no consequence in statistical hypothesis testing.[3] In fact, because our Equations 8 and 9 do include explicit reliability information, they serve well to illustrate the point that reliability does make a difference. Specifically, a decrease in test reliability must be compensated for by an increase in sample size in order to maintain equivalent statistical power. Moreover, and despite what some of Forsyth's remarks would have one believe, it is the case that *the advantage of a randomized block design relative to a completely randomized design must decrease as reliability decreases.* These points will be illustrated in the following discussion.

A re-analysis of our original example reveals the consequences of adding measurement error to initially errorless dependent and blocking variables. It should be emphasized that the arguments made here follow directly from the associated mathematical formulas and are not unique to the particular example we have devised. In the original Situation 1, both the blocking variable and the dependent variable were measured without error, i.e., $\varrho_{XX'} = \varrho_{YY'} = 1.00$; and the correlation between the two in this idealized situation was given by $\varrho_{T_X T_Y} = .50$.

If a researcher sets $\alpha = .05$ and desires power of $1 - \beta = .80$ to detect a difference between $K=2$ means of at least $\Psi_\sigma = 1.0$ true-score standard deviation unit, then a total of 17 subjects per treatment group would be required under a completely randomized design arrangement. On the other hand, were a randomized block design adopted instead, the effective standardized true-score difference increases to $\Psi^*_\sigma = \Psi_\sigma / \sqrt{1-\varrho^2_{T_X T_Y}} = 1.0/\sqrt{1-(.50)^2} = 1.155$, which requires only 14 subjects per treatment group to maintain equivalent power. Thus, the advantage of the randomized block (RB) design over the completely randomized (CR) design is realized in the savings of $K(\text{CR}-\text{RB}) = 2(17-14) = 6$ subjects in this situation.

That Forsyth's (1978) statement, ". . . it is not necessary to consider the reliabilities of the X and Y measures when comparing the relative merits of the (two designs)," is misleading is demonstrated via the original Situation 4. In this situation, measurement error has been added to both X and Y variables such that $\varrho_{XX'} = \varrho_{YY'} = .80$. Accordingly, the correlation between X and Y is attenuated as follows: $\varrho_{XY} = \sqrt{\varrho_{XX'}}\sqrt{\varrho_{YY'}}\varrho_{T_X T_Y} = \sqrt{.80}\sqrt{.80}(.50) = .40$. In the completely randomized design, our Equation 8 indicates that the measurement error associated with Y shrinks the standardized mean difference of interest from $\Psi_\sigma = 1.0$ to $\Psi_\sigma = \sqrt{\varrho_{YY'}}\Psi_\sigma = \sqrt{.80}(1.0) = .894$, thereby increasing the required number of subjects per treatment group from 17 to 21.

Similarly, if the attenuated ϱ_{XY} is inserted into our Equation 9 (as it should have been in the original article), the standardized mean difference of the randomized block design drops from $\Psi^*_\sigma = 1.155$ to

$$\Psi^*_\sigma = \sqrt{[\varrho_{XX'}\varrho_{YY'} - \varrho^2_{XY}]/[\varrho_{XX'}(1-\varrho^2_{XY})]}\,\Psi_\sigma = \sqrt{[(.80)(.80) - (.40)^2]/[.80(1-.40^2)]}(1.155) = .976.$$

The required number of subjects per treatment group is found to increase from 14 to 19. As a result, the total subject savings is 4 in Situation 4, as compared with 6 in Situation 1.

The discussion of Situations 1 and 4 adequately illustrates the two points made earlier: (1) a decrease in reliability (from 1.00 to .80) must be compensated for by an increase in sample size (from 17 to 21 for CR and from 14 to 19 for RB) in order to maintain equivalent statis-

[3]Throughout this article, as well as throughout our earlier one, it is implicit that when we speak of changes in reliability we are focusing on changes in measurement error variance (σ_E) rather than on changes in true score variance (σ_T) (for discussion of this important distinction, see Nicewander & Price, 1978).

tical power; and (2) the advantage of a randomized block design, relative to a completely randomized design, decreases as reliability decreases (from 6 to 4 total subject savings). In short, reliability does make a difference.

Two other points need to be made with regard to this example. First, it is easily shown that the subject savings advantage of RB and CR diminishes even more if lower values of $\varrho_{XX'}$ and $\varrho_{YY'}$ are selected. This implies that there do indeed exist situations in which test reliabilities are low enough so that (given other cost considerations) a researcher would be advised to adopt a completely randomized design, even though a randomized block design would prove considerably more efficient in a situation with no measurement error. This point was made in our original article. Second, from our previous Equations 10 and 11 and example, it may be noted that in a randomized block design unreliability of the dependent variable has greater sample-size compensation consequences than does unreliability of the blocking variable—assuming all else is held constant. In the present example, 18 vs. 15 subjects per randomized-block treatment group are required for Situations 2 ($\varrho_{YY'} = .80$) and 3 ($\varrho_{XX'} = .80$), respectively, assuming that the shrunken ϱ_{XY} is used in the formulas ($\Psi^*_o = 1.0$ and 1.118, respectively).

In summary, we attempted to show in our original article that

1. Errors of measurement affect (specifically, reduce) the power of the statistical test of the hypothesis,
   $H_0: \mu_1 = \mu_2 = \ldots = \mu_K$.
2. Because of this, if a researcher wishes to achieve the same statistical power in the presence of measurement error as would have been achieved in its absence, he/she will have to consider some changes.
3. Increasing the measuring instrument's reliability (e.g., by taking more measurements per subject) or increasing the number of independent sampling units (e.g., by taking more subjects per measurement), or both, are possible options.
4. Should the researcher decide on the second option and thereby increase sample size, it is of some interest to know how many additional subjects he/she would require in order to have power comparable to that in either the error-free situation or in other error-full situations which rely on alternative, typically more powerful, experimental designs (e.g., a randomized block rather than a completely randomized design).
5. Each such situation presents the researcher with a unique rational decision-making process, inasmuch as it is possible to determine the amount of power and/or sample size associated with a particular design given a priori specifications of the number of treatment groups, Type I error probability, magnitude of experimental effects expressed in true score units, instrument reliability, and the correlation between blocking and dependent variables. Because of the vast number of ways in which these specifications can vary, an experimental design that is found to be optimal in one situation may not be in another.

Having given the above summary of intent, we stand behind all of the formulas derived in the original article. We regret having made "translation" errors in applying the formulas to our example and appreciate Forsyth's (1978) pointing these out. At the same time, it would indeed be unfortunate if the general basis of our message were to become lost in Forsyth's specifics.

References

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.

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.

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.

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

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

Acknowledgments

Both authors contributed equally to this work.

Author's Address

Joel R. Levin, Department of Educational Psychology, 1025 W. Johnson St., University of Wisconsin, Madison, WI 53706.