Psychological Bulletin
1965, Vol. 64, No. 5, 326–329

CONTROL OF SPURIOUS ASSOCIATION AND THE RELIABILITY OF THE CONTROLLED VARIABLE

DANIEL KAHNEMAN [1]

Hebrew University, Jerusalem

The techniques of matched groups, analysis of covariance, and partial correlation represent various approaches to the prevention of a spurious association between X_1 and X_2 due to a confounding variable, X_3. In all these techniques the use of an unreliable measure for X_3 leads to a systematic bias of undercorrection. Adequate corrections are possible for the case of known reliability of X_3. Groups should be matched on true scores rather than observed scores, but no correction is possible for the factorial design in which groups are formed on the basis of unreliable correlated measures. Partial correlations should be corrected for the effects of unreliability of the controlled variable. Spuriously high partials are usually obtained when this correction is not applied.

Spurious correlations and confounding variables present a characteristic and recurrent problem to the social scientist. When samples of children drawn from several social classes (Variable X_1) are found to differ on the level of their professional aspirations (X_2), one asks: Are the differences due, at least in part, to social class differences in intelligence (X_3)? If a positive correlation is discovered between occupational aspirations (X_1) and college grades (X_2), the possibility that differences in attitudes toward achievement (X_3) mediate the relationship is immediately suggested.

The techniques most commonly used to remove confounding effects of X_3 on the relationship between X_1 and X_2 are various matched-group designs, the analysis of covariance, and partial correlation. Having used these techniques, the investigator is likely to feel safe: the confounding variable has been removed and need not be considered further in the interpretation of results.

The purpose of the present note is to draw attention to the effects of unreliability in the measurement of X_3 on these research designs. Such unreliability typically leads to undercorrection so that the spurious effects due to X_3 are only partly removed. Some procedures which fully correct these biases will be mentioned for the special case of known reliability of the controlled measure. No solutions are proposed to the difficulties of significance testing which arise because of the reliability issue.

[1] My thanks are due to Ester Samuel, of the Department of Statistics, for her help in viewing the problem in its proper perspective. J. Levin and M. Kubovy read the manuscript and made many helpful comments.

Matched-Group Designs and the Analysis of Covariance

The essence of the matched-group design is that cases are selected from two or more populations so that known differences among these populations in the distribution of some variable are ostensibly removed. Thus, two groups of children drawn respectively from an upper- and lower-class background may be equated on the basis of intelligence test scores. Such a matching procedure may be followed in an attempt to study the effect of social class on some dependent variable without contamination by intelligence. Thorndike (1942) has pointed out that such an approach tends to lead to spurious results when the variable which is the basis for matching is measured with less than perfect reliability. The core of the fallacy is that groups matched in this manner are similar in their observed scores but differ in their *true* scores. On retesting, the lower-class children will regress toward the mean of the population from which they have been drawn and will tend to do more poorly. The middle-class children will regress toward their population mean, and their test scores will indicate improvement.

The magnitude of the regression effect can be calculated by Formula 1, if the reliability

of the test within each population is known.

$$\bar{T} = \mu + \rho(\bar{X} - \mu), \qquad [1]$$

where $\bar{T}$ and $\bar{X}$ are, respectively, the means of true and observed scores in the group; ρ is the reliability coefficient, and μ is the mean observed score in the population.

The between-group differences in true scores of X_3 which remain after matching on observed scores make it likely that a spurious association between X_1 and X_2 will be found. The matching procedure undercorrects for such effects. Thorndike (1942) proposed that groups be matched on the basis of predicted true scores, recently termed "regressed scores" by McNemar (1962, p. 161). Formula 1 may be used for this purpose.

The caution suggested by Thorndike appears to have been largely ignored in the psychological literature, and the procedure of matching on observed scores is still commonly used. The fallacies due to unreliability and regression effects are particularly severe in a complex variant of the matched-groups design, the factorial design where the levels on the various factors are formed by selection of groups on the basis of test scores. Regression fallacies occur when measures of low reliability which are correlated in nature are made ostensibly orthogonal in the experimental population. This design is quite popular in personality research and has often been applied to measures of doubtful reliability.

A hypothetical example of this type of design is presented in Table 1. It is assumed that an investigator sets up a 2×2 factorial design on the basis of test scores on two measures, A and B. The test-retest reliabilities of these measures are .49 and .64, respectively. The correlation between them in an unselected population is .30. The table shows the mean true values to be expected on each variable when the mean observed standard scores for the high and for the low groups are .50 and $-.50$.

The values presented in Table 1 have been computed by multiple-regression equations, in which the observed values of A and B are used to predict the true scores on these variables. All scores are in standard units. The subscripts o and T refer to observed and true scores, respectively. The correlations used in the computation are: $r_{A_0B_0} = .30$, $r_{A_0A_T} = .70$, $r_{B_0B_T} = .80$, $r_{A_0B_T} = .43$, $r_{A_TB_0} = .38$. The correlations between true and observed scores are obtained by means of the appropriate corrections for attenuation. It is apparent in Table 1 that the true scores on the variables are neither matched nor orthogonal. Spurious main effects and interactions are a very likely outcome when this design is used with correlated measures of low reliability.

TABLE 1

MEAN TRUE SCORES OF GROUPS IN A FACTORIAL DESIGN FOUNDED ON OBSERVED SCORES

	Low, $\bar{A}_o = -.50$	High, $\bar{A}_o = .50$
High, $\bar{B}_o = .50$	$\bar{A}_T = -.19$ $\bar{B}_T = +.30$	$\bar{A}_T = +.43$ $\bar{B}_T = +.45$
Low, $\bar{B}_o = -.50$	$\bar{A}_T = -.43$ $\bar{B}_T = -.45$	$\bar{A}_T = +.19$ $\bar{B}_T = -.30$

It is important to note that this type of flaw is actually fatal to the design. There is no correcting formula which will simultaneously achieve the required matching of values in the rows and columns.

The analysis of covariance has been suggested as an alternative to the matching procedure which presumably overcomes the regression fallacy (McNemar, 1962, p. 373). The covariance design is generally used when there exists a substantial correlation between the dependent (X_2) and controlled (X_3) variables. Two main cases may be distinguished (Edwards, 1950) depending on whether systematic between-group differences exist on the controlled variable. In the absence of large between-group differences on X_3, unreliability in the measurement of this variable simply reduces the effectiveness of the analysis without producing any systematic bias. The case where large differences in X_3 are found is the one where matching and analysis of covariance are reasonable alternatives. Under such conditions, it can be shown that unreliability in X_3 has similar effects on the two procedures.

Consider the case where the null hypothesis is valid. When the reliability of X_3 decreases, the variance of this measure is inflated by error. Therefore, the regression of individual scores of X_2 on the observed values of X_3 within groups is of lesser slope than the corre-

sponding regression on true values of X_3. However, when real differences exist among the means of X_3 in the different populations (classes of X_1), unreliability of measurement has relatively less effect on the between-group variance of X_3: the variance of observed means is partly due to sampling error (which is affected by reliability), but it is also affected by population differences in X_3, on which reliability has no effect. Consequently, the observed regression among group means will tend to be steeper than the within-group regression, which leads to a bias in the F test. It is interesting to note that increases in sample size have no systematic effect on the expected value of within-group estimates, while such increases consistently improve the approximation of between-group estimates to the true regression. Thus, in the case described, there is an inflated probability of Type I errors over stated significance levels which is due to insufficient control of X_3. The probability of such errors increases as the reliability of X_3 decreases; it also increases with sample size.

The conclusion must be that the analysis of covariance is suspect of undercorrection whenever prior analysis suggests that real differences among groups exist on the X_3 variable, unless this variable is identified with very high reliability.

Partial Correlation

Partial correlations are used: (*a*) to establish whether an observed correlation between two variables is spurious, being due to the joint association of these two variables with a third, or (*b*) less frequently, to uncover a correlation between X_1 and X_2 which may be obscured when these variables are differently related to X_3. In both cases it can be shown that unreliability in the measurement of X_3 biases results in a systematic and predictable manner.

The partial correlation between observed values of X_1 and X_2—with *true* scores in X_3 controlled—is given in Equation 2:[2]

$$R_{12.3} = \frac{r_{12} - R_{13}R_{23}}{\sqrt{(1 - R_{13}^2)}\sqrt{(1 - R_{23}^2)}}. \quad [2]$$

[2] The symbol R is used for coefficients in which true values of X_3 are considered.

When X_3 is measured with reliability ρ, coefficients which involve this variable are attenuated. The partial correlation, with observed scores of X_3 controlled, now becomes:

$$r_{12.3} = \frac{r_{12} - \rho R_{13}R_{23}}{\sqrt{(1 - \rho R_{13}^2)}\sqrt{(1 - \rho R_{23}^2)}}. \quad [3]$$

When reliability is perfect, it can be seen that $r_{12.3}$ is equal to $R_{12.3}$. With zero reliability, $r_{12.3}$ is equal to r_{12}, that is, the measure of direct association which the use of partial correlation was intended to correct.

With intermediate values of reliability, the observed partial generally takes a value between the true partial and the direct correlation between X_1 and X_2.[3] Where the true partial is higher than the direct correlation, $r_{12.3}$ underestimates the relationship. Perhaps more important, when the true partial is substantially lower than r_{12}, the observed partial systematically overestimates the degree of unconfounded association between X_1 and X_2. In either case, the use of observed scores for X_3 in the partial coefficient yields a result which is biased in the direction of r_{12}. The covariance between X_1 and X_2 which is due to X_3 is not effectively partialled out.

The magnitude of this bias is illustrated in the following example. Consider the following correlations, in which true scores of X_3 are assumed: $r_{12} = .30$, $R_{13} = .60$, $R_{23} = .50$. The true partial correlation is 0. When X_3 is measured with a reliability of .64, the following observed correlations are expected: $r_{12} = .30$, $r_{13} = .48$, $r_{23} = .40$. The observed partial correlation is .135.

In order to obtain unbiased estimates of a partial correlation in practice, it is essential to correct r_{13} and r_{23} for attenuation due to the unreliability of X_3. Unlike the case of simple correlation, the correction for attenuation does not necessarily yield a higher estimate of the partial correlation. In the most common case, where the relationship between

[3] This statement is not always true. Under some restricted conditions, the observed partial does not vary monotonically when the reliability of X_3 increases. The conditions for this effect are closely similar to the conditions under which $R_{12.3}$ and r_{12} are identical. In such cases, partial correlation is of no practical relevance. Its value is then affected to a very limited extent by the reliability of X_3.

X_1 and X_2 is suspected of being spurious, correction for attenuation will yield a lower value for the partial correlation. In general, high values of the observed partial will be found less biased than lower values. However, when the observed partial is of borderline significance, considerations of reliability may have an important effect on research conclusions.

REFERENCES

EDWARDS, A. L. *Experimental design in psychological research.* New York: Rinehart, 1950.

MCNEMAR, Q. *Psychological statistics* (2nd ed.). New York: Wiley, 1962.

THORNDIKE, R. L. Regression fallacies in the matched groups experiment. *Psychometrika,* 1942, **7**, 85–102.

(Received August 16, 1964)