EVALUATING THE EFFECT OF INADEQUATELY MEASURED VARIABLES IN PARTIAL CORRELATION ANALYSIS

By Samuel A. Stouffer
University of Chicago

Partial and multiple correlation are used, ordinarily, in the absence of a theory as to the mathematical relationship among the variables. A simple linear combination is assumed and the principal attention is focused either on the regression equation as a predictive tool, on one of the partial correlation coefficients, or on a comparison of the so-called "relative importance" of the different independent variables.

It is not generally recognized that such an analysis assumes that each of the variables is perfectly measured, such that a second measure X'_i, of the variable measured by X_i, has a correlation of unity with X_i. If some of the measures are more accurate than others, the analysis is impaired. For example, the sociologist may have a problem in which an index of economic status and an index of nativity are independent variables. What is the effect, if the index of economic status is much less satisfactory than the index of nativity? Ordinarily, the effect will be to underestimate the significance of the less adequately measured variable and to overestimate the significance of the more adequately measured variable.

A variable may be "inadequately" measured in either or both of at least two respects: (1) The measure may have low reliability, that is, it fails to measure *something* consistently. For example, a score derived from the odd-numbered questions on a test of social attitudes may have a low correlation with the parallel score derived from the even-numbered questions on the same test. Or, the schedules in a standards-of-living study may be so badly filled out that the correlation between indexes derived from similar schedules filled out by two different interviewers of the same families may be low. (2) The measure may have high reliability, yet low validity. That is, it fails to measure adequately what it purports to measure. A reliable test may not necessarily be a valid test of social attitudes, as might be checked by correlating the test scores with some other index of social attitudes. Or, indexes derived from accurately filled out schedules in a standards-of-living study may have a low correlation with indexes from schedules based on a

different, though equally defensible, concept of standards of living.[1]

If either the reliability or validity of an index is in question, at least two measures of the variable are required to permit an evaluation. The purpose of this paper is to provide a logical basis and a simple arithmetical procedure (a) for measuring the effect of the use of two indexes, each of one or more variables, in partial and multiple correlation analysis and (b) for estimating the likely effect if two indexes, not available, could be secured.

THEORETICAL CONSIDERATIONS

Let us assume that we have s variables, of each of which there exist two measures X and X', based on n cases. Our problem is to compare the results from the use of both X_i and X'_i with the results from the use of X_i alone. The problem might be examined still more generally by considering k measures of each variable X_i. The present paper, however, is limited to a consideration of the two measures X_i and X'_i, of each of our s variables.

The writer has considered three different approaches, which, though different in their initial logic, lead, as will be proved, to identical results in important special cases.

(1) If we consider X_1, the dependent variable, satisfactorily measured, such that X'_1 may be disregarded, we may find the multiple correlation of X_1 with X_2 and X'_2 holding constant the remaining $2(s-2)$ variables. The theory was described by the writer in a previous paper in this JOURNAL.[2] Expressing v_1, v_2, and v'_2 as respective deviations from the planes

$$v_1 = X_1 - (a_1 + b_{13.3'4\ldots ss'}X_3 + b_{13'.34\ldots ss'}X'_3 + \cdots + b_{1s'.33'\ldots(s-1)'s}X'_s)$$

$$v_2 = X_2 - (a_2 + b_{23.3'4\ldots ss'}X_3 + b_{23'.34\ldots ss'}X'_3 + \cdots + b_{2s'.33'\ldots(s-1)'s}X'_s)$$

$$v'_2 = X'_2 - (a'_2 + b_{2'3.3'4\ldots ss'}X_3 + b_{2'3'.34\ldots ss'}X'_3 + \cdots + b_{2's'.33'\ldots(s-1)'s}X'_s)$$

one finds the multiple correlation between the values of v. This correlation coefficient, since it has properties both of multiple and of partial correlation, has been called $r_{1.22'.33'\ldots ss'}$, the coefficient of combined partial correlation, and it has been shown that it may be expressed in terms of conventional values of r by writing

$$r_{1.22'.33'\ldots ss'} = \sqrt{1 - (1 - r^2_{12.33'\ldots ss'})(1 - r^2_{12'.233'\ldots ss'})}. \qquad (1)$$

[1] Another type of inadequacy may arise when X_1 and X_2, say, are ratios with a common inaccurate denominator p, while X_3 does not contain p. Then r_{12} under certain conditions will be too high. (Cf. Karl Pearson, *Proceedings of the Royal Society*, lx, 1897, p. 489.) If r_{12} is too high, $r_{12.3}$ ordinarily will be overestimated as compared with $r_{12.3}$. However, this so-called "spurious correlation" is negligible in cases where it is logical to use percentages or ratios in social or demographic statistics, as G. Udny Yule has shown (*Journal of the Royal Statistical Society*, lxxiii, 1910, p. 644). The writer's present considerations do not deal with the problem discussed by Pearson and Yule and should not be confused with it.

[2] "A Coefficient of 'Combined Partial Correlation' with an Example from Sociological Data," this JOURNAL, v. 29, March, 1934, pp. 70–71.

This approach to our problem has two limitations, among others. It does not provide for two measures of the dependent variable and it requires $(2s-1)$ dimensions for handling a problem with only s sets of measures. On the other hand, if one has a problem with only three sets of measures this method provides a useful procedure for comparing $r_{1.22'.33'}$ with $r_{1.33'.22'}$ and noting how they differ from $r_{12.3}$ and $r_{13.2}$, respectively. Equation 1 may be written

$$r_{1.22'.33'} = \sqrt{\frac{r^2_{12.33'} + r^2_{12'.33'} - 2r_{12.33'}r_{12'.33'}r_{22'.33'}}{1 - r^2_{22'.33'}}}. \tag{1a}$$

If $r_{12}=r_{12'}$, and $r_{23}=r_{2'3}=r_{23'}=r_{2'3'}$, while $r_{22'}$ and $r_{33'}$ are each $\neq \pm 1$, Equation 1a reduces to a simple form in terms of zero-order r's, namely,

$$r_{1.22'.33'} = \frac{r_{12}d_{33} - r_{13}r_{23}}{\sqrt{(d_{33} - r^2_{13})(d_{22}d_{33} - r^2_{23})}}, \tag{1b}$$

where $d_{ii}=\frac{1}{2}(1+r_{ii'})$. If we possess only one index of the variable measured by X_3, Equation 1a reduces to

$$r_{1.22'.3} = \sqrt{\frac{r^2_{12.3} + r^2_{12'.3} - 2r_{12.3}r_{12'.3}r_{22'.3}}{1 - r^2_{22'.3}}}, \tag{1c}$$

which, if $r_{12}=r_{12'}$, and $r_{23}=r_{2'3}$, while $r_{22'}\neq \pm 1$, reduces to the very convenient form

$$r_{1.22'.3} = \lambda r_{12.3}, \tag{1d}$$

where $\lambda=\sqrt{(1-r^2_{23})/(d_{22}-r^2_{23})}$ and $d_{22}=\frac{1}{2}(1+r_{22'})$. It is evident from an inspection of the expression under the radical that $r_{1.22'.3}>r_{12.3}$, as would follow from the property of $r_{1.22'.3}$ as a multiple correlation coefficient. The value of $r_{1.22'.3}$ may be compared with the conventional value of $r_{13.22'}$ and it can then be noted how they differ from $r_{12.3}$ and $r_{13.2}$, respectively.

(2) Let us now avail ourselves of X_1 and X'_1, two measures of the dependent variable, and join a fourth equation to the three considered above, namely,

$$v'_1 = X'_1 - (a'_1 + b_{1'3.3'4\ldots ss'}X_3 + b_{1'3'.34\ldots ss'}X'_3 + \cdots + b_{1's'.33'\ldots(s-1)'s}X'_s).$$

Write $y_i=v_i/\sigma_{v_i}$ and form the sums $(y_1+y'_1)$ and $(y_2+y'_2)$. Since $\sigma_{y_i}=1$, whence

$$\sigma_{y_i+y'_i} = \sqrt{\sigma^2_{y_i} + 2r\sigma_{y_i}\sigma_{y'_i} + \sigma^2_{y'_i}} = \sqrt{2(1 + r_{y_iy'_i})}$$

and since $\Sigma y_iy_{i'}/n=r_{y_iy'_i}$, we have

$$r_{(y_1+y'_1)(y_2+y'_2)} = \frac{\Sigma(y_1 + y'_1)(y_2 + y'_2)}{n\sigma_{y_1+y'_1}\sigma_{y_2+y'_2}} = \frac{r_{y_1y_2} + r_{y_1y'_2} + r_{y'_1y_2} + r_{y'_1y'_2}}{2\sqrt{(1 + r_{y_1y'_1})(1 + r_{y_2y'_2})}}$$

whence, from the relation $r_{y_iy_j} = r_{ij.33'\ldots ss'}$,

$$r_{(y_1+y'_1)(y_2+y'_2)} = \frac{r_{12.33'\ldots ss} + r_{12'.33'\ldots ss'} + r_{1'2.33'\ldots ss'} + r_{1'2'.33'\ldots ss'}}{2\sqrt{(1 + r_{11'.33'\ldots ss'})(1 + r_{22'.33'\ldots ss'})}}. \quad (2)$$

It is interesting to consider again the case of three variables. Equation 2 reduces to

$$r_{(y_1+y'_1)(y_2+y'_2)} = \frac{r_{12.33'} + r_{12'.33'} + r_{1'2.33'} + r_{1'2'.33'}}{2\sqrt{(1 + r_{11'.33'})(1 + r_{22'.33'})}}. \quad (2a)$$

If we now assume that $r_{ij} = r_{ij'} = r_{i'j} = r_{i'j'}$, while $r_{ii'} \neq 1$, Equation 2a may be shown to reduce to

$$r_{(y_1+y'_1)(y_2+y'_2)} = \frac{r_{12}d_{33} - r_{13}r_{23}}{\sqrt{(d_{11}d_{33} - r^2_{13})(d_{22}d_{33} - r^2_{23})}} \quad (2b)$$

where $d_{ii} = \frac{1}{2}(1+r_{ii'})$. If we use only one index of the independent variable, Equation 2b reduces to a form identical with (1b). Moreover, if we use only one index each of X_1 and X_3, Equation 2b reduces to

$$r_{y_1(y_2+y'_2)} = \lambda r_{12.3}, \quad (2c)$$

which is identical with (1d).

It will be observed that, although $(y_1+y'_1)$ and $(y_2+y'_2)$ are index numbers formed by combining two measures of each factor, it is not necessary arithmetically to go through the process described, since Equation 2 expresses our results directly in terms of correlation coefficients between the original measures of X. The principal limitation of this approach seems to be the fact that it requires $2s$ dimensions to handle a problem involving only s pairs of measures.

(3) By our third approach we reduce the problem to s dimensions. Writing $z_i = (X_i - \overline{X}_i)/\sigma_i$, where $\overline{X} = \Sigma X_i/n$, we form the sums

$$\begin{aligned} t_1 &= z_1 + z'_1 \\ t_2 &= z_2 + z'_2 \\ &\cdots \\ t_s &= z_s + z'_s \end{aligned}$$

and seek to express the relationships among the values of t in terms of relationships among the original values of X.

Remembering that $\sigma_{z_i} = 1$, whence $\sigma_{t_i} = \sqrt{2(1+r_{ii'})}$, and that $\Sigma z_i z_j/n = r_{ij}$, where r_{ij} is the zero-order correlation between X_i and X_j, we have

$$\begin{aligned} r_{t_it_j} &= \frac{\Sigma(z_i + z'_i)(z_j + z'_j)}{2n\sqrt{(1 + r_{ii'})(1 + r_{jj'})}} \\ &= \frac{r_{ij} + r_{i'j} + r_{ij'} + r_{i'j'}}{4\sqrt{[(1 + r_{ii'})/2][(1 + v_{jj'})/2]}} = \frac{\bar{r}_{ij}}{\sqrt{d_{ii}d_{jj}}}, \end{aligned} \quad (3)$$

where $\bar{r}_{ij}=\frac{1}{4}(r_{ij}+r_{i'j}+r_{ij'}+r_{i'j'})$, the average zero-order intercorrelations, and where $d_{ii}=\frac{1}{2}(1+r_{ii'})$.

Consider now the conventional correlation matrix

$$\Delta = \begin{vmatrix} 1 & r_{t_1t_2} & \cdots & r_{t_1t_s} \\ r_{t_1t_2} & 1 & \cdots & r_{t_2t_s} \\ \cdot & \cdot & \cdots & \cdot \\ r_{t_1t_s} & r_{t_2t_s} & \cdots & 1 \end{vmatrix}, \tag{3a}$$

which, upon substitution of the values of $r_{t_it_j}$ found in (3), becomes

$$\Delta = \begin{vmatrix} 1 & \dfrac{\bar{r}_{12}}{\sqrt{d_{11}d_{12}}} & \cdots & \dfrac{\bar{r}_{1s}}{\sqrt{d_{11}d_{ss}}} \\ \dfrac{\bar{r}_{12}}{\sqrt{d_{11}d_{22}}} & 1 & \cdots & \dfrac{\bar{r}_{2s}}{\sqrt{d_{22}d_{ss}}} \\ \cdot & \cdot & \cdots & \cdot \\ \dfrac{\bar{r}_{1s}}{\sqrt{d_{11}d_{ss}}} & \dfrac{\bar{r}_{2s}}{\sqrt{d_{22}d_{ss}}} & \cdots & 1 \end{vmatrix}. \tag{3b}$$

Multiply the elements in the first row by $\sqrt{d_{11}}$, the elements in the second row by $\sqrt{d_{22}}$, etc. Multiply the elements in the first column by $\sqrt{d_{11}}$, the elements in the second column by $\sqrt{d_{22}}$, etc. We have

$$\Delta = \frac{1}{d_{11}d_{22}\cdots d_{ss}} \begin{vmatrix} d_{11} & \bar{r}_{12} & \cdots & \bar{r}_{1s} \\ \bar{r}_{12} & d_{22} & \cdots & \bar{r}_{2s} \\ \cdot & \cdot & \cdots & \cdot \\ \bar{r}_{1s} & \bar{r}_{2s} & \cdots & d_{ss} \end{vmatrix} \tag{3c}$$

and we write

$$\Delta = \frac{\Delta'}{d_{11}d_{22}\cdots d_{ss}}.$$

As a consequence of the operation in passing from (3b) to (3c), any $(s-1)$-rowed minor of Δ',

$$\Delta'_{ij} = \frac{\sqrt{d_{ii}d_{jj}}}{d_{11}d_{22}\cdots d_{ss}}\Delta_{ij}, \tag{3d}$$

where Δ'_{ij} is formed by crossing out the i'th row and j'th column in Δ'. This is a special case, where $m_i=2$, of a more general determinant in which $d_{ii}=[1+(m_i-1)\bar{r}_{ii}]/m_i$, in which $\bar{r}_{ii}$ is the average of the $m_i(m_i-1)/2$ intercorrelations between m_i measures of a given varia-

ble i. It is of interest to note that as $m_i \to \infty$, $d_{ii} \to \bar{r}_{ii}$, whence, if m_j also $\to \infty$, $\bar{r}_{ij}/\sqrt{d_{ii}d_{jj}} \to \bar{r}_{ij}/\sqrt{\bar{r}_{ii}\bar{r}_{jj}}$, which is a form of the correlation coefficient corrected for attenuation.[3]

Solution of this determinant gives the equations needed, in terms of zero-order correlation coefficients, for the complete analysis of our problem. Equation 3c makes explicit the assumption, as to the intrinsic accuracy of all variables, which is implicit in the conventional partial correlation analysis. Only as every value $r_{ii'} \to 1$, whence $d_{ii} = \frac{1}{2}(1+r_{ii'}) \to 1$, and as every value $\bar{r}_{ij} \to r_{ij}$, does Δ' approach the usual form. In other words, the customary correlation analysis assumes that every X_i would correlate perfectly with another measure X'_i, of the same variable.

Let us now consider three variables only, namely, t_1, t_2, and t_3. Equation 3c becomes

$$\Delta = \frac{1}{d_{11}d_{22}d_{33}} \begin{vmatrix} d_{11} & \bar{r}_{12} & \bar{r}_{13} \\ \bar{r}_{12} & d_{22} & \bar{r}_{23} \\ \bar{r}_{13} & \bar{r}_{23} & d_{33} \end{vmatrix}. \tag{3e}$$

The four types of values in which there is likely to be most interest with respect to our present problem are $r_{t_1t_2 \cdot t_3}$, $\beta_{t_1t_2 \cdot t_3}$, $\beta_{t_1t_2 \cdot t_3}r_{t_1t_2}$, and $R^2_{t_1 \cdot t_2t_3}$. In the interest of clarity, the notation will be changed by writing $(i+i)$ in the place of t_i in the subscripts. Let us now express our desired measures in terms of the zero-order correlations between the original values of X_1, X_2, and X_3, remembering that $\bar{r}_{ij} = \frac{1}{4}(r_{ij}+r_{ij'}+r_{i'j}+r_{i'j'})$, and that $d_{ii} = \frac{1}{2}(1+r_{ii'})$.

(a) *When there are two measures each of* X_1, X_2, *and* X_3.

$$\begin{aligned} r_{(1+1)(2+2)\cdot(3+3)} &= \frac{\Delta_{12}}{\sqrt{\Delta_{22}\Delta_{11}}} = \frac{\Delta'_{12}}{\sqrt{\Delta'_{22}\Delta'_{11}}}, \quad \text{from (3d)}, \\ &= \frac{\bar{r}_{12}d_{33} - \bar{r}_{13}\bar{r}_{23}}{\sqrt{(d_{11}d_{33} - \bar{r}^2_{13})(d_{22}d_{33} - \bar{r}^2_{23})}}. \end{aligned} \tag{3f}$$

It will be observed that if we write $r_{ij} = \bar{r}_{ij}$, (3f) becomes identical with (2b); otherwise, (3f) may be expected to differ from (2) because of the different logic behind the respective derivations.

$$\begin{aligned} \beta_{(1+1)(2+2)\cdot(3+3)} &= \frac{\Delta_{12}}{\Delta_{11}} = , \qquad \text{from (3d)}, \\ \frac{\Delta'_{12}}{\Delta'_{11}}\sqrt{\frac{d_{22}}{d_{11}}} &= \left(\frac{\bar{r}_{12}d_{33} - \bar{r}_{13}\bar{r}_{23}}{d_{22}d_{33} - \bar{r}^2_{23}}\right)\sqrt{\frac{d_{22}}{d_{11}}}. \end{aligned} \tag{3g}$$

[3] For the general proof see the writer's paper, "Reliability Coefficients in a Correlation Matrix," *Psychometrica*, June, 1936. Equation 3 can be shown to be a special case of Equation 147 in Truman L. Kelley, *Statistical Methods*, p. 197.

$$\beta_{(1+1)(2+2).(3+3)}r_{(1+1)(2+2)} = \left[\frac{\bar{r}_{12}d_{33} - \bar{r}_{13}\bar{r}_{23}}{d_{11}(d_{22}d_{33} - \bar{r}^2_{23})}\right]\bar{r}_{12}. \tag{3h}$$

$$\begin{aligned} R^2_{(1+1).(2+2)(3+3)} &= 1 - \frac{\Delta}{\Delta_{11}} = 1 - \frac{\Delta'}{\Delta'_{11}d_{11}} \\ &= \frac{\bar{r}^2_{12}d_{33} + \bar{r}^2_{13}d_{22} - 2\bar{r}_{12}\bar{r}_{13}\bar{r}_{23}}{d_{11}(d_{22}d_{33} - \bar{r}^2_{23})}. \end{aligned} \tag{3i}$$

(b) *When there are two measures each of X_1, and X_2, and when there is one measure of X_3.* Substitute $d_{33}=1$ in Equation 3e or Equations 3f to 3i, inclusive. Example:

$$r_{(1+1)(2+2).3} = \frac{\bar{r}_{12} - \bar{r}_{13}\bar{r}_{23}}{\sqrt{(d_{11} - \bar{r}^2_{13})(d_{11}d_{22} - \bar{r}^2_{23})}}. \tag{3j}$$

If we write $r_{ij}=\bar{r}_{ij}$, (3j) becomes identical with results obtained from reducing (2b).

(c) *When there is one measure of X_1 and when there are two measures each of X_2 and X_3.* Substitute $d_{11}=1$ in (3e) or Equations 3f to 3i, ininclusive. Example:

$$r_{1(2+2).(3+3)} = \frac{\bar{r}_{12}d_{33} - \bar{r}_{13}\bar{r}_{23}}{\sqrt{(d_{33} - \bar{r}^2_{13})(d_{22}d_{33} - \bar{r}^2_{23})}}. \tag{3k}$$

If we write $r_{ij}=\bar{r}_{ij}$, Equation 3k becomes identical with (1b), or with (2b) when $d_{11}=1$. The three approaches to our problem coincide in results at this point, although, if $r_{ij}\neq\bar{r}_{ij}$, we may expect differences.

(d) *When there is one measure each of X_1, and X_3, while there are two measures of X_2.* Substitute $d_{11}=d_{33}=1$ in Equation 3e or Equations 3f to 3i, inclusive. Write $r_{13}=\bar{r}_{13}$. Examples:

$$r_{1(2+2).3} = \frac{\bar{r}_{12} - \bar{r}_{13}\bar{r}_{23}}{\sqrt{(1 - r^2_{13})(d_{22} - \bar{r}^2_{23})}}, \tag{3l}$$

which, if $r_{12}=r_{12'}$ and if $r_{23}=r_{2'3}$, reduces, exactly as (1b) and (2b) reduce, to

$$r_{1(2+2).3} = \lambda r_{12.3}, \tag{3m}$$

where $\lambda=\sqrt{(1-r^2_{23})/(d_{22}-r^2_{23})}$, an identity with (1d) or (2c). Moreover,

$$r_{13.(2+2)} = \frac{r_{13}d_{22} - \bar{r}_{12}\bar{r}_{23}}{\sqrt{(1 - r^2_{13})(d_{22} - \bar{r}^2_{23})}}. \tag{3n}$$

If $r_{12}=r_{12'}$ and if $r_{23}=r_{2'3}$, we may write (3n) as

$$r_{13.(2+2)} = \frac{r_{13}d_{22} - r_{12}r_{23}}{\sqrt{(1 - r^2_{13})(d_{22} - r^2_{23})}}, \tag{3o}$$

which can be shown to be identical with the conventional formula for $r_{13.22'}$ under the same assumptions, where $r_{13.22'}$ is the partial correlation coefficient between X_1 and X_3, with X_2 and X'_2 held constant. When $r_{12}=r_{12'}$ and $r_{23}=r_{2'3}$ and when $r_{12.3}$, $r_{13.2}$, $r_{1(2+2).3}$, and $r_{12.(3+3)}$ each $\neq 0$, we write

$$\frac{r_{1(2+2).3}}{r_{13.(2+2)}} = k\frac{r_{12.3}}{r_{13.2}}, \qquad \text{whence}$$

$$k = \frac{\lambda r_{13.2}}{r_{13.(2+2)}} = \frac{r_{13} - r_{12}r_{23}}{r_{13}d_{22} - r_{12}r_{23}} > 1,$$

if r_{13} is positive and $>r_{12}r_{23}$, or if r_{13} is negative and $<r_{12}r_{23}$.

If $r_{12}\neq r_{12'}$ and if $r_{23}\neq r_{2'3}$, the logic of our derivation would require that $r_{1(2+2).3}$ be compared with $r_{13.(2+2)}$, rather than with $r_{13.22'}$. It will be observed that the arithmetical operations needed to calculate (3n) are simpler than those needed to calculate $r_{13.22'}$.

By similar methods the reader may find easily the values of $r_{(1+1).23}$ or of any other functions derived from the correlation matrix.

We have seen that when $r_{ij}=r_{ij'}=r_{i'j}=r_{i'j'}$, our second and third theoretical approaches lead to identical values of partial r, and that when $r_{11'}=1$, where X_1 is the dependent variable, our first approach also coincides in results. It is the writer's judgment that the third approach is to be preferred, both theoretically and practically, because of its simplicty and generality. It reduces a problem with $2s$ sets of measures to one of s dimensions. It permits a ready comparison not only of such values as $r_{1(2+2).3}$ and $r_{13.(2+2)}$, or some functions thereof, but also of such values as $\beta_{1(2+2).3}$ and $\beta_{13.(2+2)}$, or the products of the Betas with $r_{1(2+2)}$ and r_{13}, respectively, or of such a value as $R^2_{1.(2+2)3}$. It avoids logical difficulties as to dependent and independent variables which might possibly appear from the application of least square theory in the second approach, and it permits the computation of standard errors by conventional formulas. Each of the three approaches assumes X_i and X'_i to be of equal weight or value for use in an index.

The third approach, it will be remembered, assumes that an index number t_i is formed by finding $z_i=(X_i-\overline{X}_i)/\sigma_i$ and $z'_i=(X'_i-\overline{X}'_i)/\sigma'_i$ and adding these two standard measures. It is possible, especially if X is a fraction and X' is another measure of $(1-X)$, that X and X' will be negatively correlated. Naturally, in combining X and X' in an

index, a research worker would reverse the signs either of z or z', making the correlation positive. This is not strictly required in the theoretical development above; except that if $r_{ii'}$ is negative the problem becomes indeterminate when $r_{ii'} = -1$. Arithmetically, of course, it is not necessary to compute the index number t_i, as Equations 3f to 3i, inclusive, or any other measures derived from Equation 3c, may be computed directly from the correlation coefficients involving the original measures of X and X', taken individually. If $r_{ii'}$ is negative, one should change the sign of $r_{ii'}$ to positive and reverse the signs in all other correlation coefficients involving X'.

Finally, it often happens that one has some reason to believe that a particular index is inadequate, yet has no second measure at hand. Nevertheless, he would like to know roughly how much difference it might make in his final interpretation if some second index could have been used. If he is willing to assume that the correlations of his unknown second index with the other variables would be the same as the correlations of his known first index with these variables, he can set an upper and lower limit of discrepancy by arbitrarily assigning to the unknown $r_{ii'}$ a low value and then a high value. In the special case where Equation 3m is applicable, no computation is required, as values of λ in (3m) are presented in Table I for selected values of $r_{22'}$ and r_{32}, or, rather, more generally for selected values of $r_{jj'}$ and r_{jk}. It should be said with emphasis, however, that values derived by making these assumptions never should be reported *in lieu of* $r_{12.3}$ or $r_{13.2}$. The new

TABLE I

VALUES OF $\lambda = \sqrt{(1-r^2_{jk})/(d_{jj}-r^2_{jk})}$ FOR USE IN THE EQUATION $r_{i(j+j').k} = \lambda r_{ij.k}$
[Assuming that $r_{ij} = r_{ij'}$, and that $r_{jk} = r_{j'k}$, and writing $d_{jj} = \frac{1}{2}(1 + r_{jj'})$]

r_{jk}	$r_{jj'} = +.50$	$r_{jj'} = +.60$	$r_{jj'} = +.70$	$r_{jj'} = +.80$	$r_{jj'} = +.90$
.00	1.155	1.118	1.085	1.054	1.026
.05	1.155	1.118	1.085	1.054	1.026
.10	1.157	1.119	1.086	1.055	1.026
.15	1.159	1.121	1.087	1.055	1.027
.20	1.163	1.124	1.089	1.057	1.027
.25	1.168	1.127	1.091	1.058	1.028
.30	1.174	1.132	1.094	1.060	1.029
.35	1.183	1.138	1.098	1.062	1.030
.40	1.193	1.146	1.103	1.065	1.031
.45	1.207	1.155	1.110	1.069	1.033
.50	—	1.168	1.118	1.074	1.035
.55	—	1.184	1.129	1.080	1.038
.60	—	—	1.143	1.089	1.042
.65	—	—	1.162	1.100	1.046
.70	—	—	—	1.115	1.053
.75	—	—	—	1.139	1.063
.80	—	—	—	—	1.078
.85	—	—	—	—	1.104

($r_{i(j+j').k} = \lambda r_{ij.k}$) is Equation 3m in this paper.

values are supplements to the information obtained from $r_{12.3}$ and $r_{13.2}$, not substitutes, and may be used cautiously as guides only. The same caution, of course, does not apply to the use of the more general results when all of the zero-order correlations are known, although in any case, the limitation must be kept in mind that X_i and X'_i are receiving equal weights.

ILLUSTRATIONS OF THE APPLICATION

(1) Suppose that we are interested in the question, "Why do residents of some areas of a large city move their abodes less often than residents of other areas?" We should guess that stability of residence must be closely related to home ownership. We also should guess that stability may be related to the presence of larger than average families who have a good many young children.

Using 1934 data for 651 Chicago census tracts,[4] we have three indexes:

X_1 = percentage of families residing at their present abode at least five years prior to the 1934 census.
X_2 = percentage of families with four or more members.
X_3 = percentage of families owning their own homes.

We take X_1 as an index of stability in an area, X_2 as an index of larger than average families, and X_3 as an index of home ownership. Finding $r_{12}=.6475$, $r_{13}=.8501$, and $r_{23}=.6055$, we obtain $r_{12.3}=.317$ and $r_{13.2}=.755$.

Unfortunately, our index of larger than average families is unsatisfactory, because it fails to measure adequately the variable in which we are really interested, namely, the presence of larger than average families who have a good many young children. That is, we are questioning the validity of the index when it is to be used as an index of what we want to measure, because it fails to discriminate between families which may be composed wholly of adults and families which are composed partly of small children. It happens that we know the ratio of children under 5 to women 20 to 44 in each tract. Let us call this ratio X'_2 and introduce it as a fourth variable in a conventional correlation analysis. Since $r_{12'}=.5158$, $r_{22'}=.6646$, and $r_{2'3}=.4283$, we have $r_{12.2'3}=.175$, $r_{12'.23}=.179$, and $r_{13.22'}=.758$. Evidently, both of our family indexes now almost vanish as compared with our index of home ownership. But a moment's reflection will indicate that in the present case $r_{12.2'3}$ and $r_{12'.23}$ have little, if any, realistic meaning.

[4] The data, including the zero-order correlation coefficients, were generously supplied by Richard O. Lang, fellow in sociology at the University of Chicago. The writer also is indebted to Mr. Lang for assistance in computation, especially in the preparation of Table I.

What we are really interested in is the combined association of X_2 and X'_2 with X_1, as compared with the association of X_3 with X_1.

We decide to form a new family index, $t_2 = z_2 + z'_2$, where $z_2 = (X_2 - \overline{X}_2)/\sigma_2$ and $z'_2 = (X'_2 - \overline{X}'_2)/\sigma'_2$. The computation of this index would be laborious, however, as there are 651 tracts. We can save the labor and get identical results by simply using our observed zero-order correlation coefficients in Equation 31 of the present paper. The computation takes practically no more time than that leading to first-order partials and, of course, much less time than that leading to second-order partials such as those in the preceding paragraph. We find that $r_{1(2+2).3}$, the correlation between the index of stability and the new and more inclusive family index, holding constant the index of home ownership, is .396, by Equation 31, while $r_{13.(2+2)}$, the correlation between stability and home ownership, holding constant the new family index, is .776, by Equation 3n. We see that $r_{1(2+2).3}$ is about twenty per cent larger than $r_{12.3}$, while $r_{13.(2+2)}$ (which, in most problems, would be smaller than $r_{13.2}$) is only two per cent larger than $r_{13.2}$.

We have been assuming that our index of stability and our index of home ownership are satisfactory. We recall, however, from a study using 1930 census tract data in Cleveland, Ohio[5] that a correlation of only .85 was found between X_3, the percentage of families owning their own homes and X'_3, the percentage of homes owned per 100 dwellings. A reason for the discrepancy is that if an area contains only two-family dwellings, the maximum home ownership by our index could be only 50 per cent, or if an area contains only four-family dwellings the maximum home ownership could be only 25 per cent. For our Chicago series no values of X'_3 have been computed, though they might be obtained if necessary. In the Cleveland study, p. 217, we see that X_3 and X'_3 correlated about alike with several other social and economic variables, none of which, however, correspond to our X_1, X_2, or X'_2. Assuming that the correlations of X'_3 with X_1, X_2, and X'_2 would be about the same as the respective correlations of X_3 with these variables, and assuming that for Chicago $r_{33'}$ would be .90 at the minimum, because we have observed that $r_{13} = .85$, we can estimate what our results might have been if X'_3 had been combined with X_3 in a new index of home ownership. Little additional computation is required. Setting $d_{11} = 1$, $d_{22} = \frac{1}{2}(1 + r_{22'}) = .8323$, $d_{33} = \frac{1}{2}(1 + r_{33'}) = .95$, $\bar{r}_{12} = \frac{1}{2}(r_{12} + r_{12'}) = .58165$, $r_{13} = .8501$, and $\bar{r}_{23} = \frac{1}{2}(r_{23} + r_{2'3}) = .5169$, we we substitute in Equation 3k of the present paper, obtaining

[5] Henry D. Sheldon, Jr., "Problems in the Statistical Study of Juvenile Delinquency," *Metron*, xii, December, 1934, pp. 201–23.

$r_{1\,(2+2)\,.\,(3+3)} = .328$, while $r_{1\,(3+3)\,.\,(2+2)} = .800$ is obtained after interchanging transcripts 2 and 3 in the same formula. Thus, the inclusion of a second index of home ownership, provided our assumptions hold, may lower $r_{1\,(2+2)\,.3}$ about 17 per cent and raise $r_{13\,.\,(2+2)}$ about 3 per cent. On the basis of this information, we can decide whether or not it is worth while to work up the actual data for X'_3 and bring X'_3 into the problem formally. We might, indeed, decide to neglect both X'_2 and X'_3, since our last result is closer to the original than the second. But we now have information to guide us in our decision.

For comparative purposes, the values discussed, together with some additional values which may be of interest, are recorded below. (Incidentally, the independent computation of the square of the multiple correlation coefficient by two different formulas may be used, as in the conventional correlation analysis, as an automatic check on the arithmetic used in calculating the partial r's and β's.)

$r_{12.3}$	$=.317$	$r_{1(2+2).3}$	$=.396$	$r_{1(2+2).(3+3)}$	$=.328$, estimated.
$r_{13.2}$	$=.755$	$r_{13.(2+2)}$	$=.776$	$r_{1(3+3).(2+2)}$	$=.800$, estimated.
$\beta_{12.3}r_{12}$	$=.136$	$\beta_{1(2+2).3}r_{1(2+2)}$	$=.166$	$\beta_{1(2+2).(3+3)}r_{1(2+2)}$	$=.126$, estimated.
$\beta_{13.2}r_{13}$	$=.615$	$\beta_{13.(2+2)}r_{13}$	$=.612$	$\beta_{1(3+3).(2+2)}r_{1(3+3)}$	$=.661$, estimated.
$R^2_{1.23}$	$=.751$	$R^2_{1.(2+2)3}$	$=.778$	$R^2_{1.(2+2)(3+3)}$	$=.787$, estimated.

(2) Let us suppose that in the foregoing problem we had reason to feel satisfied with X_1 and X_2. Our information from the Cleveland study leads us to wonder how much our values of $r_{12\,.3}$ and $r_{13\,.2}$ would be altered if we improved the index X_3 by combining with it X'_3. Assuming that $r_{13'}$ would equal r_{13} and that $r_{23'}$ would equal r_{23}, and writing $r_{33'} = .90$ on the same grounds as in the second paragraph preceding, we have, from Equation 3m,

$$r_{1(3+3)\,.2} = \lambda r_{13.2}$$

where λ may be found without computation, simply by entering our Table I, with $r_{jk} = r_{32} = .85$ and $r_{jj'} = r_{33'} = .90$. We see that $\lambda = 1.104$, and therefore estimate $r_{1\,(3+3)\,.2} = 1.104 \times .755 = .83$. To estimate $r_{12\,.\,(3+3)}$ on the same assumptions, we need only to substitute our observed r_{12}, r_{13}, and r_{23} and our guessed value of $d_{33} = \frac{1}{2}(1 + r_{33'}) = .95$ in Equation 3o (after an interchange of transcripts in 3o), obtaining $r_{12\,.\,(3+3)} = .28$, which is about 10 per cent less than $r_{12\,.3} = .317$.

(3) Returning again to the Cleveland study, we use a different set of data. We seek the relationship between X_1, the juvenile delinquency rate in 1928–31 by census tracts, X_2, an index of dependency in 1928,

and X_3, the percentage of native whites in the population. Given $r_{12}=.75$, $r_{13}=-.51$, and $r_{23}=.60$, from p. 206, we have $r_{12.3}=.65$. After the study is completed, a parallel index for 1931 becomes available. Call it X'_2. Shall we include it in the study? Assume that we have no knowledge of $r_{12'}$, $r_{22'}$, and $r_{2'3}$. Since $r_{12}=.75$, we are probably justified in assuming that $r_{22'}$ is at least .80. While the dependency rate in 1931 is higher throughout the city than in 1928, we have no *a priori* reason to assume that the relationships between dependency and delinquency and nativity have changed markedly. Entering Table I with $r_{jk}=r_{23}=.60$ and $r_{jj}=r_{22'}=.80$, we find $\lambda=1.089$. Hence, we estimate by Equation 3m, $r_{1(2+2).3}=1.089\times.65=.71$, and conclude that with the use of a more reliable index of dependency $r_{1(2+2).3}$ will lie somewhere between .65 (which is $r_{12.3}$) and .71. In this case, actual data happen to be available, p. 218, namely, $r_{12'}=.77$, $r_{22'}=.90$, and $r_{2'3}=.64$, permitting us to use Equation 3l, from which we calculate $r_{1(2+2).3}=.69$.

It is hoped that this paper will interest research workers sufficiently to encourage further exploration of the theoretical approaches here examined. Further empirical study of the range of safety in the use of the approximation formulas also is desirable. From the standpoint of application if there is a hesitance, because of the time required, to use these or better methods which subsequent students may develop, one can say only that an extra few minutes spent in analyzing one's correlation problem is a trivial amount of time as compared with the time taken to collect or reduce the data.