Psychological Bulletin
1976, Vol. 83, No. 6, 1007-1013

Comparison of Correlations, Variances, Covariances, and Regression Weights With or Without Measurement Error

C. E. Werts and D. A. Rock
Educational Testing Service
Princeton, New Jersey

R. L. Linn
University of California, Los Angeles

K. G. Jöreskog
University of Uppsala, Sweden

A maximum likelihood procedure for testing the equality of sets of variances, covariances, correlations, and regression weights between and/or within populations is demonstrated. This procedure is an application of Jöreskog's general factor-analytic model for simultaneous factor analysis in several populations.

Jöreskog (1971) devised SIFASP, a general computer program (van Thillo & Jöreskog, Note 1) for simultaneous maximum likelihood factor analysis in several populations. This paper shows how SIFASP can be used to test the equality of correlations, variances, covariances, and regression weights within and/or between samples of different sizes. Corrections for measurement error can be readily incorporated.

Program Description

SIFASP assumes that a factor analysis model holds in each of the g populations under study. If $\mathbf{X}_g$ is defined as the vector of the p observed measures in group g, then $\mathbf{X}_g$ can be accounted for by k common factors (f_g) and p unique factors (z_g). In matrix terms, the model in each population is:

$$\mathbf{X}_g = \boldsymbol{\mu}_g + \Lambda_g f_g + z_g \,, \qquad (1)$$

where μ_g is a vector of means and Λ_g is a matrix of factor loadings for the p observed scores on the k common factors in the gth population. The p unique factors are independent of each other and of the common factors. Without loss of generality it may be assumed that the expected value of f_g and z_g is zero, that is, that the common and unique factors have zero means.

Requests for reprints should be sent to Charles E. Werts, Developmental Research Division, Educational Testing Service, Princeton, New Jersey 08540.

Given these factor model assumptions, the expected variance-covariance matrix Σ_g among the observed scores has the form:

$$\Sigma_g = \Lambda_g \Phi_g \Lambda_g' + \Psi_g^2 \,, \qquad (2)$$

where Φ_g is the variance-covariance matrix of f_g and Ψ_g^2 is the diagonal variance-covariance matrix of z_g. When the factor model does not fit the data perfectly, the observed variance-covariance matrix S_g for the gth group will differ from Σ_g. The computer program yields a chi-square statistic that is a measure of how much Σ_g differs from S_g, that is, of how well the model fits the data. It is not permissible to standardize the variables in each group and to analyze the correlation matrices instead of the variance-covariance matrices. This violates the likelihood function, which is based on the distribution of the observed variances and covariances.

The three matrices Λ_g, Φ_g, and Ψ_g^2 for each group are called the *pattern* matrices. The elements of the pattern matrices are the model parameters, and parameters are of three kinds: (a) *fixed* parameters, which have been assigned given values, like 0 or 1; (b) *constrained* parameters, which are unknown but equal to one or more other parameters; and (c) *free* parameters, which are unknown and not constrained to be equal to any other parameter. A parameter may be constrained to be equal to other parameters in the same and/or different pattern matrices in the same and/or in different groups.

TABLE 1

THE VARIANCE–COVARIANCE MATRIX FOR GROUP 1

Item	X_1	X_2	X_3	X_4
X_1 = a 40-item verbal aptitude section	63.382			
X_2 = a separately timed 50-item verbal aptitude section	70.984	110.237		
X_3 = a 35-item math aptitude section	41.710	52.747	60.854	
X_4 = a separately timed 25-item math aptitude section	30.218	37.489	36.392	32.295

TESTING THE EQUALITY OF VARIANCE–COVARIANCE MATRICES BETWEEN POPULATIONS

The logic of this test is to create a set of factors that are identical to the observed measures, which means that the variance–covariance matrix Φ_g among the factors is identical to the matrix among the observed measures. Constraining the corresponding elements of Φ_g across groups, therefore, is a test of the equality of the variance–covariance matrices between groups. The factors identical to the observed measures are created by making the factor-loading matrix Λ_g an identity matrix (i.e., unities in the diagonal, zeros elsewhere) and Ψ_g^2 a null matrix (i.e., all zeros), indicating no residual variances. This setup is equivalent to the Box (1949) test.

Example

Tables 1 and 2 give the observed variance–covariance matrices for two random samples ($N_1 = 865$, $N_2 = 900$, respectively) of candidates who took the Scholastic Aptitude Test (Swineford, Note 2) in January 1971. The four measures are, in order: X_1 = a 40-item verbal aptitude section, X_2 = a separately timed 50-item verbal aptitude section, X_3 = a 35-item math aptitude section, and X_4 = a separately timed 25-item math aptitude section.

The factor-loading matrices Λ_1 and Λ_2 are 4×4 identity matrices, the residual matrix $\Psi_1^2 = \Psi_2^2 = 0$, and Φ_1 and Φ_2 are 4×4 variance–covariance matrices in which corresponding elements have been constrained equal. Using SIFASP, the maximum likelihood estimate of $\Phi = \Phi_1 = \Phi_2$ is:

$$\hat{\Phi} = \begin{bmatrix} 65.687 & & & \\ 71.702 & 108.753 & & \\ 41.120 & 54.074 & 62.054 & \\ 29.586 & 38.207 & 37.857 & 33.881 \end{bmatrix}.$$

The discrepancies between $\hat{\Phi}$ and the observed matrices are calculated by the program $(\hat{\Phi} - S_g)$:

Population 1 residuals

$$= \begin{bmatrix} 2.305 & & & \\ 0.718 & -1.484 & & \\ -0.590 & 1.327 & 1.200 & \\ -0.632 & 0.718 & 1.465 & 1.586 \end{bmatrix}$$

and

Population 2 residuals

$$= \begin{bmatrix} -2.211 & & & \\ -0.689 & 1.423 & & \\ 0.571 & -1.273 & -1.149 & \\ 0.610 & -0.689 & -1.404 & -1.522 \end{bmatrix}.$$

This model yielded a chi-square of 32.85 with 10 degrees of freedom (corresponding to the 10 elements set equal). The probability of ob-

TABLE 2

THE VARIANCE–COVARIANCE MATRIX FOR GROUP 2

Item	X_1	X_2	X_3	X_4
X_1 = a 40-item verbal aptitude section	67.898			
X_2 = a separately timed 50-item verbal aptitude section	72.301	107.330		
X_3 = a 35-item math aptitude section	40.549	55.347	63.203	
X_4 = a separately timed 25-item math aptitude section	28.976	38.896	39.261	35.403

taining a larger chi-square when the model of equal variance–covariance matrices is true is $p = .000$. It may be concluded that the equal variance-covariance matrix hypothesis is inconsistent with the data, since the chi-square is highly significant (at less than the .001 level). Whether this statistically significant difference is of practical importance can be better judged from the absolute magnitude of the discrepancies noted above. We judge these discrepancies to be of little practical importance.

A major advantage over the Box test is that the elements of Φ_g can be tested for equality across and or within populations one or more at a time. Thus, a chi-square fit statistic could have been obtained for each element, testing its equality across groups. A subset of elements that are equal can be located in this way and tested as a subset for equality across groups.

Testing the Equality of Correlation Matrices Between Populations

The logic of this test is to create a set of factors in each group with unit variances, which means that Φ_g for each group is a correlation matrix. Constraining the corresponding elements of Φ_g across groups, therefore, is a test of the equality of the correlation matrices between groups. This is accomplished by fixing the diagonal elements of Φ_g at unity and specifying the off-diagonal elements to be free. The factor-loading matrix Λ_g is a diagonal matrix with the diagonal elements free and other elements fixed at zero. There are no residuals (Ψ_g^2 is all zeros), since in essence the observed variables are simply being rescaled into factors with unit variance. The corresponding off-diagonal elements of Φ_g are constrained across populations. The degrees of freedom for the chi-square fit test are equal to the number of unique correlations in Φ_g.

Example

The data in Table 1 were used to test the hypothesis that the correlation matrices were equal. The program yielded the following factor-loading estimates:

$$\hat{\Lambda}_1 = \begin{bmatrix} 7.916 & .0 & .0 & .0 \\ .0 & 10.487 & .0 & .0 \\ .0 & .0 & 7.816 & .0 \\ .0 & .0 & .0 & 5.680 \end{bmatrix}$$

and

$$\hat{\Lambda}_2 = \begin{bmatrix} 8.286 & .0 & .0 & .0 \\ .0 & 10.372 & .0 & .0 \\ .0 & .0 & 7.935 & .0 \\ .0 & .0 & .0 & 5.953 \end{bmatrix}.$$

The hypothesized population correlation matrix was estimated as:

$$\hat{\Phi} = \begin{bmatrix} 1.000 & & & \\ .849 & 1.000 & & \\ .645 & .658 & 1.000 & \\ .629 & .630 & .826 & 1.000 \end{bmatrix}.$$

The program uses these parameter estimates to generate an estimated variance–covariance matrix $\hat{\Sigma}$ for each group using Equation 2 and reports discrepancies from the observed variance–covariance matrices. It is of interest to calculate the discrepancies between $\hat{\Phi}$ and the observed correlations. Deleting diagonal elements, $[\hat{\Phi} - S_g]$ is:

$$\text{Population 1} \begin{bmatrix} .000 & & \\ -.027 & .014 & \\ -.039 & .002 & .005 \end{bmatrix}$$

and

$$\text{Population 2} \begin{bmatrix} .001 & & \\ .026 & -.014 & \\ .038 & -.001 & -.004 \end{bmatrix}.$$

It is common practice to obtain the root mean square discrepancy by taking the square root of the average of all squared discrepancies. A value of .020 was obtained for this example. This can be compared with the chi-square of 24.20 ($df = 6$), which is significant at less than the .001 level. Even though there is a statistically significant difference, the root mean square discrepancy of .020 (largest value .038) is not ordinarily of practical importance.

Analysts who find it easier to interpret discrepancies between correlations could have converted $\hat{\Phi}$ in the Example section under the heading Testing the Equality of Variance–Covariance Matrices Between Populations to a correlation matrix and calculated differences from the observed correlation matrices.

Comparisons Within Populations

In this section we demonstrate the flexibility of SIFASP with respect to rescaling; in effect testing a ratio of one parameter to another.

Consider again the example used in the section entitled Testing the Equality of Variance–Covariance Matrices Between Populations. The observed variances of the two verbal sections should differ because of unequal lengths (likewise with the math sections). There should, however, be a specifiable relationship because these sections were intended to be "parallel" tests with different numbers of items. To find the expected relationship between the observed variances, the principles detailed by Lord and Novick (1968, p. 86) were used to derive the following equation:

$$\frac{\text{var}\ (X_2)}{\text{var}\ (X_1)} = \frac{m(m+k)}{(1+k)},$$

where $k = (m+1)(1-R_c)/R_c$, R_c = reliability of the composite test $X_1 + X_2$, and m = ratio of number of items of X_2 to X_1.

The Group 2 data in Table 2 were used in the published analysis of the January 1971 Scholastic Aptitude Test (Swineford, Note 2). The published whole test reliability is $R_c = .928$, from which a k of .175 was calculated. With $m = 50/40 = 1.25$, the ratio of the observed variances should be 1.516 if the sections are parallel.

To test a ratio of variances, the variance of X_2 was rescaled to reduce its variance by 1.516. This is done by setting the factor loading to $\sqrt{1.516} = 1.231$. The factor-loading matrix is therefore fixed at:

$$\Lambda_2 = \begin{bmatrix} 1.0 & .0 & .0 & .0 \\ .0 & 1.231 & .0 & .0 \\ .0 & .0 & 1.0 & .0 \\ .0 & .0 & .0 & 1.0 \end{bmatrix}.$$

The residuals ($\Psi_2{}^2$) were again set equal to zero, and in Φ the first two diagonal elements were set equal. A chi-square of 1.43 with 1 degree of freedom was obtained. Since the probability of obtaining a larger chi-square if the model is true is $p = .23$, it was concluded that this model is consistent with the data. Swineford's (Note 2) statement that the two sections may be regarded as essentially parallel is consistent with these findings. It could also have been shown that if the sections are parallel, then the relation of the variance of X_1 to the covariance between X_1 and X_2 should be

$$\text{var}(X_1) = [(1+k)/m]\ \text{cov}(X_1X_2).$$

By appropriate rescaling, this hypothesis could be checked.

In part, this example was chosen to demonstrate that the factor loadings can be changed in any meaningful way. An additional element of flexibility across populations is that the rank of the observed matrices need not be the same in the different groups—for example, the correlation matrix of one group may be compared with the appropriate subset of a larger matrix in another group. If the comparable variables are ordered differently in the observed matrices, it is only necessary to make sure that the comparable parameter elements are equated.

Testing Regression Weights

Consider the case of the regression of X_2 on X_1 in which it is hypothesized that this regression weight (B) is the same in two groups. The logic in this case is to set up a factor (f_1) identical to the independent variable, in which case the factor loading of X_2 on this factor will be identical to the regression weight for X_2 on X_1. Constraining the corresponding factor loadings across groups tests the equality of the respective regression weights. Because there is a residual in the regression of X_2 on f_1, this problem requires the use of Ψ^2, which for convenience is defined as the diagonal of the variance–covariance matrix of the independent residuals.

The equations corresponding to Equation 1 are:

$$X_1 = f_1$$

and

$$X_2 = Bf_1 + \zeta\,.$$

The factor-loading matrix in Equation 1 is therefore:

$$\Lambda_g = \begin{bmatrix} 1 \\ B \end{bmatrix}.$$

Since there is only one factor, Φ_g will consist of a single element, the variance of f_1 that is free:

$$\Phi_g = [\text{var}(f_1)]\,.$$

Since there is no residual for the X_1 on f_1 regression, the first residual variance will be fixed at zero and the second residual variance will be the variance of the regression residuals, $\text{var}(\zeta)$:

$$\text{diagonal } \Psi^2 = [0, \text{var}(\zeta)],$$

where $\text{var}(\zeta)$ is a free parameter. By constraining the B_g across groups, it is possible to test the equality hypothesis without simultaneously testing the equality of the regression residual (ζ_g) variances. The latter may be tested separately or in combination with the equal B_g assumption. This procedure is equivalent to that of McLaughlin (1975).

Testing Partial Regression Weights

In the general case it is necessary to define a factor (f_k) for each independent variable (X_k) such that $X_k = f_k$. The loadings of the dependent variable on these factors are the partial regression weights that may be tested for equality between and/or within groups. For example, if X_1, X_2, and X_3 are the independent variables and X_4 the dependent variable, then

$$\Lambda_g = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ B_{41} & B_{42} & B_{43} \end{bmatrix},$$

and

$$\Phi_g = \begin{bmatrix} \text{var}(f_1) & & \\ \text{cov}(f_1 f_2) & \text{var}(f_2) & \\ \text{cov}(f_1 f_3) & \text{cov}(f_2 f_3) & \text{var}(f_3) \end{bmatrix},$$

and

$$\text{diagonal } \Psi_g^2 = [0, 0, 0, \text{var}(\zeta)].$$

It is quite possible to test the equality of regression weights between two groups in which the number of independent variables differs. Another practical application of this technique is the testing of the homogeneity of within-groups regressions when covariates are used in repeated measures designs. This is a special case of the procedures outlined in this section or the section entitled Testing Regression Weights, depending on whether there are single or multiple covariates.

Standardized Partial Regression Coefficients

A somewhat more interesting problem arises when standardized weights are to be compared. The difficulty is that the factors for the independent variables may be readily standardized but the dependent variable is an observed variable that cannot be standardized without violating the likelihood function assumptions. The answer arises from the fact that the factors may be assigned any arbitrary variance without restricting the likelihood function. The chi-square remains the same because changing the factor variance changes the factor loading inversely by the same amount, that is, the procedure is a simple rescaling that does not affect fit. Consider what would happen in a model that had only the regression of X_2 on X_1, where $X_1 = \lambda_1 f_1$ and $X_2 = Bf_2 + \zeta$. If the variance of f_1 is fixed at the variance of X_2, then $\lambda_1 = \sqrt{\text{var}(X_1)/\text{var}(X_2)}$ and B = correlation of X_1 and X_2 or the standardized regression coefficient of X_2 on X_1. If B were tested for equality across groups, it would be the equivalent of testing the equality of the correlation between X_1 and X_2.

In the general case, if each of the independent variable factors has its variance fixed at that of the dependent variable, then the covariance of each factor with the dependent variable is standardized. The covariance between each pair of factors is equal to the correlation between the corresponding pair of independent variables. It follows that the loadings of the dependent variable on the factors are standardized partial regression weights. A test of the equality of the residual variance across groups is equivalent to testing whether the coefficient of multiple correlation is equal. Any regression coefficient can be tested for equality to a fixed value of zero by setting it equal to zero and examining the significance of the resulting chi-square. This test is equivalent to testing whether the corresponding partial, part, or semi-partial correlation is different from zero.

Models When Measurement Error Is Known

Psychometricians use a variety of procedures to estimate the reliability of a test. Knowing the reliability (R_{kk}) of a test (X_k) is equivalent to knowing the error variance, since the variance of the errors is equal to the variance of the test multiplied by $(1 - R_{kk})$. Such estimates may be entered into all the models discussed previously.

Correcting Variance–Covariance Matrices

If error variances were known, then in the sections entitled Testing the Equality of Variance–Covariance Matrices Between Populations and Testing the Equality of Correlation Matrices Between Populations, Ψ_g^2 would not be set equal to zero. For the example in the variance–covaraince matrices section, diagonal

$$\Psi_g^2 = [\text{var}(E_1),\ \text{var}(E_2),\ \text{var}(E_3),\ \text{var}(E_4)],$$

where $\text{var}(E_k)$ are variances of the corresponding errors (E_k). If the error variances for both groups are entered into Ψ_1^2 and Ψ_2^2 and Λ_g and Φ_g are specified as in the variance–covariance matrices section, the resulting model would test whether the variance–covariance matrices, corrected for attenuation, were equal. The covariances between the factors are equal to the covariances between the corresponding X_k regardless of measurement error if the model is true.

When error variances are entered in Ψ_g^2 for all groups given the correlation matrices section specifications for Λ_g and Φ_g, the hypothesis is that the correlation matrices between groups are equal when corrected for attenuation.

When standard errors are entered into Ψ_2^2, for the example in Comparisons Within Populations, the hypothesis to be tested is that the variance of f_2 is $m^2 = (1.25)^2$ times the variance of f_1 and the error variance for X_2 is $m = 1.25$ times that for X_1. The variance of f_2 could be rescaled by setting the factor loading equal to 1.25, to test whether the factor variances are in the expected ratio.

Correcting Regression Weights for Attenuation

Consider the model in the section entitled Testing Regression Weights when the error variance for X_1 (i.e., $\text{var}(E_1)$) is known. In this case $\Psi_g^2 = [\text{var}(E_1), \text{var}(\zeta)]$ and the variance of f_1 will be less than the variance of X_1 by the error variance. If there are measurement errors in the dependent variable, then what is labeled $\text{var}(\zeta)$ will be the sum of the regression residuals (ζ) and the error. The estimate of B is only biased by errors in the independent variable. If the error variance for X_2 is known, then the unattenuated regression residual variance can be obtained by subtraction from the estimated variance of (ζ).

The case of partial regression weights when the error variances are known is much the same. The model in Testing Partial Regression Weights is identical except that Ψ_g^2 has one free and three fixed elements, $[\text{var}(E_1), \text{var}(E_2), \text{var}(E_3), \text{var}(\zeta)]$, where $\text{var}(\zeta)$ will include measurement error. Unless error estimates are included for all independent variables, none of the partial weights will, in general, be fully corrected for attenuation. If a partial weight corrected for attenuation is significantly different from zero, then the corresponding partial, part, or semi-partial correlation is also different from zero with the same sign.

Correcting Standardized Regression Weights for Attenuation

Consider the two-variable regression X_2 on X_1 when the error variances are known for X_1 and X_2. The model is $X_1 = \lambda_1 f_1 + E_1$ and $X_2 = Bf_1 + \zeta + E_2$. If the variance of F_1 is fixed at the variance of X_2 minus the variance of F^2, then

$$\lambda_1 = \sqrt{[\text{var}(X_1) - \text{var}(E_1)]/[\text{var}(X_2) - \text{var}(E_2)]}$$

and B = correlation of X_1 and X_2 divided by

$$\sqrt{(\text{reliability of } X_1)(\text{reliability of } X_2)}.$$

The unattenuated regression weight is equal to the correlation between X_1 and X_2 corrected for attenuation in X_1 and X_2. In general, standardized regression weights should be estimated from the unattenuated correlations among variables. In the case of multiple independent variables, the variance of the dependent variable minus its error variance is assigned to each of the independent variable factors.

Discussion

Levy (1975) provided a procedure for comparing correlations and variances that requires independent samples of equal size. In contrast, the SIFASP technique allows for comparisons within and/or between as many as 10 groups of possibly unequal size.

Lord (1975) has also provided a maximum likelihood procedure that can be used to compare (within the between samples) parameters that are a specifiable function of the observed variance–covariance matrices, such as variances, covariances, correlations, and regression weights. Comparison of the Lord and Jöreskog (1971) procedures is beyond the scope of this paper.

REFERENCE NOTES

1. Van Thillo, M., & Jöreskog, K. G. *SIFASP—A general computer program for simultaneous factor analysis in several populations* (Research Bull. 70–62). Princeton, N.J.: Educational Testing Service, 1970.
2. Swineford, F. *Test analysis, College Entrance Examination Board, January 1971, Scholastic Aptitude Test TSA13.* Princeton, N.J.: Educational Testing Service, 1971.

REFERENCES

Box, G. E. P. A general distribution theory for a class of likelihood criteria. *Biometrika*, 1949, *36*, 317–346.

Jöreskog, K. G. Simultaneous factor analysis in several populations. *Psychometrika*, 1971, *36*, 409–426.

Levy, K. J. Large-sample many–one comparisons involving correlations, proportions, or variances. *Psychological Bulletin*, 1975, *82*, 177–179.

Lord, F. M. Automated hypothesis tests and standard errors for nonstandard problems. *The American Statistician*, 1975, *29*, 56–59.

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

McLaughlin, D. H. A test for homogeneity of regression without homogeneity of variance. *Educational and Psychological Measurement*, 1975, *35*, 79–86.

(Received May 9, 1975)