A NOTE ON TESTING FOR CONSTANT RELIABILITY IN REPEATED MEASUREMENT STUDIES

Alvin J. Silk*

Working Paper 1007-78 July 1978

A NOTE ON TESTING FOR CONSTANT RELIABILITY IN REPEATED MEASUREMENT STUDIES

Alvin J. Silk*

Working Paper 1007-78 July 1978

*Professor of Management Science
Sloan School of Management
Massachusetts Institute of Technology

ABSTRACT

This paper discusses the potential usefulness of applying tests for the equality of variances (and covariances) to data from repeated measurement studies prior to estimating reliability components and coefficients. In situations where only two rounds of repeated measures are available, a test for the equality of the two (correlated) variances affords a means of checking the consistency of data with a condition necessary for a test-retest correlation to have a straightforward interpretation as a reliability coefficient. In cases where more than two waves of observations have been obtained, a test of the hypothesis that all the variances are equal and all the covariances are equal provides evidence as to the possible constancy of measure reliability across several waves of observations and is therefore relevant to the selection of an appropriate method for estimating reliability. An illustrative application of the tests is presented.

Papers by Heise (1969) and Wiley and Wiley (1970) have discussed the estimation of reliability coefficients from repeated measurements when a true change occurs between adjacent measures and the usual psychometric error model of parallel measurements (Lord and Novick, 1968, pp. 41-50) does not hold. As Wiley and Wiley (1970, pp. 113) have noted, if the true change involves a simple additive shift whereby a true score at time $t + 1$ increases (or decreases) by some fixed amount (that is identical for all observations) from their previous levels at t, then the true score variance will be equal for both sets of measures. Provided the error variances are stable, two rounds of measurements are sufficient in such a circumstance to estimate reliability which would, of course, by definition be the same at t and $t + 1$. Winer (1962, pp. 124-30) has discussed ANOVA procedures for estimating reliability components from two or more waves of repeated measurements under the assumptions of a constant difference between two pairs of repeated measures and stable error variance. However, Wiley and Wiley argue that the assumption of stable true score variances across repeated measurements is implausible in "most cases of practical interest" and go on to develop a method for estimating reliability coefficients for a situation where the process of change can be modeled in terms of linear relationships between adjacent true scores and where error variance remains constant. Lord and Novick (1968, p. 218) and Coleman (1970, p. 453) had pointed out earlier that at least three waves of repeated

measurements are required to identify the reliability parameters for such a change model.

The purpose of this note is to draw attention to the potential usefulness of some available significance tests in assessing the consistency of repeated measurements data with different assumptions about the stability of variances. Prior to actually applying some method of reliability estimation to a body of data from a repeated measurement study, consideration needs to be given to what assumptions are tenable concerning the stability of true and error variances. For example, evaluation of the reliability of instruments used in sociological research is often based on test-retest studies (see, for example, Robinson et al., 1968) and so the question may arise as to whether or not it is reasonable to assume such data satisfy the condition of constant true and error variances which is necessary for a reliability coefficient to be estimated from only two sets of repeated measurements. Similarly, in the case of studies involving more than two waves of measurements, one may wish to determine if the assumption of constant reliability across repeated measurements is contradicted by the nature of the data, thereby indicating the appropriateness of employing Wiley and Wiley's estimation procedure rather than that described in Winer and referred to above. To illustrate how certain available statistical tests may be used for such diagnostic purposes, we present some

empirical results from a repeated measurements study where the assumption of constant reliability appeared to hold.

ILLUSTRATIVE APPLICATION

The data examined here were originally reported by Palda (1966, p. 18) and consist of a measure of the "awareness" stage in the adoption process model found in the sociological literature on the diffusion of innovations (Rogers and Shoemaker, 1971) and routinely used in marketing research. Awareness levels for a newly launched consumer good were monitored in each of thirty cities at three different points in time, separated by two month intervals. Thus, the matrix of observations available for analysis corresponds to a panel design involving three waves of measurements on the same sample of size thirty (cities). The awareness measurements are proportions and were derived from telephone surveys conducted with samples of four hundred respondents drawn separately in each time period in each city. Since the sampling variance of a bionomial proportion is dependent upon the mean, over time shifts in mean awareness levels would lead to heterogeneity in variances if the raw awareness proportions were used in the analyses. A suitable variance stabilizing transformation can be used to circumvent this condition. The angular or arcsin transformation has this property and was applied here, i.e., $X=\sin^{-1}(A)^{1/2}$, where X is the transformed score and A is the original awareness proportion. If all the error variance is due to sampling a bionomial proportion, the sampling variance in the arsin scale (degrees)

is equal to 820.7/n. Thus, the sampling variance of a proportion so transformed is no longer dependent upon the mean and is essentially a constant for a given size sample, n (Snedecor and Cochran, 1967, p. 325). Table 1 presents the variance-covariance matrix and some other relevant summary statistics for the three waves of awareness measurements in the arsin scale. As expected for a diffusion process, the mean awareness levels increase monotonically over time but note that the variances exhibit a nonmonotonic pattern of fluctuations.

INSERT TABLE 1 HERE

Consider first the statistics for the first two periods: the covariance, Cov (X_1,X_2)= 34.695 and the variances, Var (X_1)= 38.422, Var (X_2)= 40.313. The related value of the product moment correlation was $r(X_1,X_2)$= .882. Now suppose one were interested in determining whether this correlation could be viewed as a conventional test-retest measure of reliability. For the usual error model, a necessary condition for a test-retest correlation to represent a reliability coefficient is that the true score and error variances be constant. If the true score and error components are independent, the variance of the observed scores is simply the sum of the constant true and error variances and therefore within the limits of sampling error, one should expect to find that the observed variances for the test and retest scores are equal[1/]. This suggests a test be made of the following

null hypothesis:

$$\text{Var } (X_1) = \text{Var } (X_2)$$

Assuming the underlying distribution of the variates is bivariate normal, the test for equality of two correlated variances due to Pitman (1939) and described in Snedecor and Cochran (1967, pp. 195-197) may be used to assess the above hypothesis[2/]. Rejection of the null hypothesis here would imply that either the true score variance or the error variance (or both) was (were) not constant for both sets of observations and so their reliabilities could not be equal. Applying the aforementioned test to the above variances, Var (X_1) and Var (X_2), we find the value of the relevant t statistic to be .269 (df=28) which is clearly not significant, and the hypothesis of constant observed score variances in the first two periods appears tenable. Thus, in this case, the test provides evidence to support acceptance of $r(X_1,X_2)$ as a measure of reliability. Testing for the equality of observed score variances in studies where only two rounds of repeated measurements have been obtained can serve as a safeguard against test-retest correlations being misinterpreted as reliability coefficients when the underlying data are not a suitable basis for the assessment of reliability [3/].

A related test bearing on the question of constant reliability may be employed when data are available from more than two waves of

repeated measurements. By the same line of reasoning as that noted above for the case of test-retest observations, if the score and error variances are constant for each wave of the repeated measurements, then it follows that all observed score variances should be equal and all the covariances between the observed scores should also be equal[4/]. A consistency check for the latter conditions may be obtained for the variance and covariances shown in Table 1 by testing the following composite null hypothesis:

$$\text{Var}\ (X_1) = \text{Var}\ (X_2) = \text{Var}\ (X_3) \quad ,$$

$$\text{Cov}\ (X_1,X_2) = \text{Cov}\ (X_1,X_3) = \text{Cov}\ (X_2,X_3) \ .$$

A likelihood-ratio test for such a hypothesis under the assumption that the variates follow a multivariate normal distribution has been developed by Box (1950, pp. 372-276) and is also described in Winer (1962, pp. 370-374) and Morrison (1976, p. 250). The value of the relevant chi square statistic for these data is 1.364 (df=4) which is not significant (.90>p>.80) and the null hypothesis of equal variances and equal covariances cannot be rejected. Hence the notion that all three waves of measurements have the same reliability can be maintained and the analysis of variance method suggested by Winer (1962, pp. 124-130) can be applied to estimate the reliability components. In

the present context, the ANOVA model is given by:

$$X_{it} = U + C_i + R_t + \varepsilon_{it}$$

where X_{it} is the arcsin transformed value of the awareness score for the city i in time period t (i=1,...,30 and t=1,2,3); U is the grand mean; C_i is the effect of city i; R_t is the effect of the t^{th} time period; and ε_{it} is the random error component.

Table 2 summarizes the ANOVA results. We observe that the F statistic for the "between time periods" effect is highly significant indicating that an additive shift occurred in the mean awareness level in at least one of the time periods. Note that value of the error variance is estimated to be 5.848 [5/]. It is interesting to note that applying Wiley and Wiley's model of linearly related adjacent true scores to the present data, yields an estimated error variance of 4.59 which is somewhat smaller than that obtained above under the assumption of a simple additive shift in true scores. Using the ANOVA components from Table 2 in Winer's recommended computational procedure, we obtain an estimate of .849 for the reliability of a single awareness measure.

INSERT TABLE 2 HERE

SUMMARY

This note has discussed the possible value of applying tests for equality of variances (and covariances) to data from repeated measurements studies prior to estimating reliability components and coefficients from them. In situations where only two rounds of repeated measures are available, a test for the equality of correlated variances affords a means of checking the consistency of data with a condition necessary for a test-retest correlation to have a straightforward interpretation as a reliability coefficient. In cases when more than two waves of repeated observations have been obtained, a test of the hypothesis that all the variances are equal and all the covariances are equal bears on the question of whether or not the reliability of the measure employed can be considered constant across the several waves of observations and hence is relevant to the selection of an appropriate method for estimating reliability. The normality assumption underlying the tests illustrated does, of course, represent a restriction on their applicability.

TABLE 1

SUMMARY STATISTICS FOR THREE WAVES OF AWARENESS MEASUREMENTS*

(Arcsin Transformation)

	X_1	X_2	X_3
X_1	38.422	.882	.831
X_2	34.695	40.313	.835
X_3	31.659	32.597	37.783
Mean	25.164	26.531	28.643

*The entries on the main diagonal are variances. Covariances are below the main diagonal and the product moment correlation coefficients are above the main diagonal. The last row contains the means for each of the three waves of measurements.

TABLE 2

ANOVA SUMMARY

Source of Variation	Mean Square	d.f.	F-ratio
Between Cities	104.825	29	
Within Cities	8.729	60	
Between Time Periods	(92.297)	(2)	(15.783*)
Error	(5.848)	(58)	
TOTAL	40.04	89	

*p<.01

FOOTNOTES

1. That is, the observed score (X) is assumed to be the sum of a true score (τ) and an error component (ε), $X = \tau + \varepsilon$. With repeated measurements, it is further assumed that errors are serially uncorrelated.

2. If the true scores and errors were normally distributed, then the observed scores would also be normally distributed.

3. It is worth noting however, that if the inequality of observed score variances for two waves of measurement is generated by the change model proposed by Wiley and Wiley, a test-retest correlation does have a reliability interpretation, albeit not a straightforward one. Assuming as Wiley and Wiley do that (1) the true score on the second test is linearily related to that for the first test and (2) the error variances are constant, then one can easily show that the product moment correlation between the two sets of measures is equal to the geometric mean of the separate reliability coefficients for the two rounds of measurements. Also, the square of such a test-retest correlation would provide a lower bound for the two unobserved reliabilities because the values of all three of these quantities must lie between zero and one. See Silk (1977) for a discussion of these points.

4. Recall that for the usual error model, the covariance between repeated measures is equal to their (constant) true score variance.

5. This may be compared to an expected value of 820.7/400=2.05 if all the error variance was due to sampling a binomial process with samples of four hundred respondents as were used in the surveys for each time period and city from which these awareness measures were obtained. Thus, a substantial amount of non-sampling error appears to be present here.

REFERENCES

Box, G.E.P.
1950 "Problems in the Analysis of Growth and Wear Curves." Biometrics 6:362- 89.

Coleman, J.S.
1968 "The mathematical study of change." Pages 428-278 in H.M. Blalock and A.B. Blalock (eds.), Methodology in Social Research. New York: McGraw-Hill.

Heise, D.R.
1969 "Separating reliability and stability in test-retest correlation." American Sociological Review 34:43-101.

Lord, F.M. and Novick, M.R.
1968 Statistical Theories of Mental Test Scores. Reading, Massachusetts: Addison-Wesley.

Morrison, D.F.
1976 Multivariate Statistical Methods. 2nd edition. New York: McGraw-Hill.

Palda, K.S.
1966 "The hypothesis of a hierarchy of effects: a partial evaluation." Journal of Marketing Research 3:13-24.

Pitman, E.J.G.
1939 "A note on normal correlation." Biometrika 31:9-12.

Robinson, J.P., Rusk, J.G., and Head, K.B.
1968 Measures of Political Attitudes. Ann Arbor, Michigan: Institute of Social Research, University of Michigan.

Rogers, E.M. and Shoemaker, F.F.
1971 Communication of Innovations. New York: Free Press.

Silk, A.J.
1977 "Test-retest correlations and the reliability of copy tests." Journal of Marketing Research 14:476-486.

Snedecor, G.W. and Cochran, W.G.
1967 Statistical Methods. 6th ed. Ames, Iowa: Iowa State University Press.

Wiley, D.E. and Wiley, J.A.
1970 "The estimation of measurement error in panel data." American Sociological Review 35:112-117.

Winer, B.J.
1962 Statistical Principles In Experimental Design. New York: McGraw-Hill.