Using Longitudinal Data to Estimate Reliability

Henk Blok and Wim E. Saris
University of Amsterdam

Werts, Breland, Grandy, and Rock (1980) have analyzed the relationship between a direct and an indirect measure of writing ability. Werts et al. assumed that the same true score underlies both measures and concluded that the test-retest reliability of the essay tests is biased due to correlated errors. The present analysis of their data shows that the direct and indirect tests measure two different abilities which correlate only .89 with each other and that it is not necessary to include correlated measurement errors for the essay tests. It is argued that the assumption that different tests measure the same ability should always be tested. Werts et al. (1980) did not test this assumption, and their conclusions, as a result, are incorrect.

The test-retest correlation is a common reliability estimate. The use of this estimate is not without problems. Test-retest correlations can lead to biased reliability estimates (1) when there is instability in the true scores in the interval between tests or (2) when the measurement errors are correlated. Werts, Breland, Grandy, and Rock (1980) have attempted to show how longitudinal data can be used to estimate reliability in the presence of correlated measurement errors. Their basic data consisted of the correlations for the Test of Standard Written English (TSWE) measured at three points in time and the ratings of essays administered at approximately the same times. The tests had been given to 234 college freshmen three times during the academic year. The correlation matrix from their study is presented in Table 1.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 7, No. 3, Summer 1983, pp. 295–301

0146-6216/83/030295-07$1.60

Werts et al. (1980), specified a model for these data by adding the three following assumptions to the usual assumption of local independence of the observed scores:

1. At each testing period the TSWE score and the essay rating represent the same true score.
2. The essay ratings may have correlated errors through time.
3. The third true score is dependent only on the second true score, and not on the first.

This model can be specified as follows. Define τ_1, τ_2, and τ_3 as the true scores underlying TSWE and the essay test at each of the three testing periods. Then, in the standardized case

$$\begin{aligned} x_1^* &= \lambda_1^*\tau_1^* + \varepsilon_1^* \\ x_2^* &= \lambda_2^*\tau_2^* + \varepsilon_2^* \\ x_3^* &= \lambda_3^*\tau_3^* + \varepsilon_3^* \\ x_4^* &= \lambda_4^*\tau_1^* + \varepsilon_4^* \\ x_5^* &= \lambda_5^*\tau_2^* + \varepsilon_5^* \\ x_6^* &= \lambda_6^{P}\tau_3^{P} + \varepsilon_6^* \end{aligned} \qquad [1]$$

where the λ^*'s are regression weights, and errors are assumed to be uncorrelated with the true scores. According to the second assumption, the errors for the essay test (ε_4^*, ε_5^*, ε_6^*) can be correlated. The analysis of Werts et al. (1980) seemed to show that the correlation between the three true scores (τ_1^*, τ_2^*, τ_3^*) is unity, a specific case of the third assumption. In addition, the regression weights within each of the two sets of measures were approximately equal.

Table 1
Data of Werts et al. (1980) N = 234

Variable	Mean	S.D.	Intercorrelations					
x_1	43.55	10.83	1.000					
x_2	46.61	9.91	.837	1.000				
x_3	48.30	10.01	.854	.842	1.000			
x_4	6.69	2.18	.621	.640	.602	1.000		
x_5	7.02	2.26	.602	.636	.551	.564	1.000	
x_6	7.32	2.53	.596	.617	.597	.572	.523	1.000

After the model was simplified according to the results found by Werts et al. (1980), the analysis of their final model was repeated. The results are presented in Table 2 (Model A). These are within rounding errors of the results presented by Werts et al. With a chi-square of 13.26 with 16 degrees of freedom, it is clear that the model fits the data. Leaving out the correlated errors leads to a chi-square of 36.5 with 17 degrees of freedom. It therefore seems necessary to introduce these correlated errors for the essay ratings.

Be that as it may, it will be argued here that the three assumptions made by Werts et al. (1980) are not necessary and can be tested. There is no reason why the second assumption, concerning the correlated errors of the essay ratings, could not apply to the TSWE scores. This assumption has therefore been changed accordingly, and the results are also presented in Table 2 (Model B). It is clear that this model fits the data just as well as the model of Werts et al., but the results are much less in favor of the TSWE measures. A choice between the two models can therefore not be made.

The present authors believe that the source of this ambiguity is the unjustified first assumption, that the TSWE and the essay test have the same true score. The TSWE measures the ability of students to find mistakes in sentences and is essentially an indirect test of writing ability. In contrast, the essay test is a direct measure of writing ability. Although there can be little doubt that the direct and the indirect tests of writing ability are closely related to each other, correlation coefficients between them are not unity (Breland & Gaynor, 1979). Therefore, a more general model has to be formulated which allows a test of the first assumption.

Alternative Models

In this section alternative models will be presented. None of these models requires either perfect correlation between the true scores or correlated errors. These two assumptions made by Werts et al. (1980) are not necessary, and it was therefore decided not to introduce them. If x_1 to x_6 represent the observed scores, τ_1 to τ_6 the true scores, and ε_1 to ε_6 the measurement errors, according to classical test theory,

$$x_i = \tau_i + \varepsilon_i. \qquad [2]$$

where

$$E(x_i) = E(\tau_i) = E(\varepsilon_i) = 0 \quad \text{for all } i;$$
$$E(\tau_i\varepsilon_j) = 0 \quad \text{for all } i, j;$$
$$E(\varepsilon_i\varepsilon_j) = 0 \quad \text{for all } i \neq j.$$

In this case x_1 to x_3 represent the TSWE measures at three points in time, and x_4 to x_6 represent the essay measures at the same three points in time. If $E(\varepsilon_i\varepsilon_i)$ is denoted by θ_{ii}, the error variances can be written as θ_{11} to θ_{66}. It is further assumed, as suggested by Werts et al., that

$$\begin{aligned} \tau_2 &= \beta_{21}\tau_1 + \zeta_2 \\ \tau_3 &= \beta_{32}\tau_2 + \zeta_3 \\ \tau_5 &= \beta_{54}\tau_4 + \zeta_5 \\ \tau_6 &= \beta_{65}\tau_5 + \zeta_6 \end{aligned} \qquad [3]$$

while

$$E(\zeta_i) = 0 \quad \text{for all } i;$$
$$E(\zeta_i\zeta_j) = 0 \quad \text{for all } i \neq j;$$
$$E(\zeta_i\tau_j) = 0 \quad \text{for all } i, j.$$

Table 2
Standardized Maximum Likelihood Estimates for the Model with Correlated Essay Errors (Model A) and the Model with Correlated TSWE Errors (Model B)

Parameters	Model A	Model B
$\lambda_1^* = \lambda_2^* = \lambda_3^*$	.919	.816
$\lambda_4^* = \lambda_5^* = \lambda_6^*$	.676	.744
$\rho(\varepsilon_1^*, \varepsilon_2^*) = \rho(\varepsilon_2^*, \varepsilon_3^*) = \rho(\varepsilon_1^*, \varepsilon_3^*)$	0[a]	.178
$\rho(\varepsilon_4^*, \varepsilon_5^*) = \rho(\varepsilon_5^*, \varepsilon_6^*) = \rho(\varepsilon_4^*, \varepsilon_6^*)$	.207	0[a]
Goodness-of-fit statistic	$\chi^2_{16} = 13.26$	$\chi^2_{16} = 13.26$

[a]These parameter values were specified by hypothesis.

If $E(\zeta_i\zeta_i)$ is denoted by Ψ_{ii}, the variances of the disturbance terms can be written as Ψ_{22}, Ψ_{33} (for the TSWE measures) and Ψ_{55}, Ψ_{66} (for the essay tests). In this formulation, the covariance between the true scores for the first TSWE and essay measures is represented by $\sigma_{\tau_4\tau_1}$. The assumption of Werts et al. (1980) that the TSWE and the essay tests represent the same true score means that $\sigma_{\tau_4\tau_1}$ is restricted to $(\sigma_{\tau_4\tau_4}\sigma_{\tau_1\tau_1})^{1/2}$. This assumption will not be made here. Instead, $\sigma_{\tau_4\tau_1}$ is treated as a free parameter which will be estimated. It will be clear from this formulation that only the third assumption of Werts et al. has been made, viz., that the two sets of true scores represent a simplex structure. The present authors' alternative model is the most general system which can be formulated for this case.

Unfortunately, the system is not identified because it is impossible to separate the variances of the disturbance terms (Ψ_{ii}) from the variances of the measurement errors (θ_{ii}). As a result, the corresponding parameters cannot be uniquely estimated without further restrictions. There are three different restrictions or constraints which can be imposed, alone or in combination. This leads to eight different forms of test equivalence. The three restrictions are

1. Constraints on the error variances: $\theta_{11} = \theta_{22} = \theta_{33}$ and $\theta_{44} = \theta_{55} = \theta_{66}$. This means that the error variances of the two tests remain stable over time. (This is the assumption suggested by Wiley and Wiley, 1970.)
2. Fixing the four β's to 1.0. The meaning of this is that the measurement scale remains the same over time.
3. Restricting the variance of the disturbance terms to zero: $\Psi_{22} = \Psi_{33} = \Psi_{55} = \Psi_{66} = 0$. When this condition is fulfilled, the correlations between the true scores within the two sets of three measures are unity and each set of three measures can be seen as congeneric instruments.

Comparisons Among the Models

The eight different models of test equivalence can now be examined in some detail, from the most restricted to the most general.

1. $\beta_{21} = \beta_{32} = \beta_{54} = \beta_{65} = 1$, $\theta_{11} = \theta_{22} = \theta_{33}$, $\theta_{44} = \theta_{55} = \theta_{66}$, $\Psi_{22} = \Psi_{33} = \Psi_{55} = \Psi_{66} = 0$. This is the most restricted model. Essentially, it is the model for parallel tests, implying that the true scores and the variances remain the same over time. As a consequence, the reliabilities remain unchanged.

2. $\beta_{21} = \beta_{32} = \beta_{54} = \beta_{65} = 1$, the θ's unconstrained, $\Psi_{22} = \Psi_{33} = \Psi_{55} = \Psi_{66} = 0$. This model implies that the true score variances remain the same, while the error variances are allowed to change. The model is known as the model for tau-equivalance tests.
3. The four β's unconstrained, $\theta_{11} = \theta_{22} = \theta_{33}$, $\theta_{44} = \theta_{55} = \theta_{66}$, $\Psi_{22} = \Psi_{33} = \Psi_{55} = \Psi_{66} = 0$. The implication of this model is the reverse of Model 2. The error variances remain the same, while the true score variances may change. (This model is suggested by Wiley and Wiley, 1970).
4. Both the four β's and the θ's unconstrained, $\Psi_{22} = \Psi_{33} = \Psi_{55} = \Psi_{66} = 0$. While the true score and the error variances are allowed to vary, the true scores are still linearly related. This model is known as the congeneric test model.
5. $\beta_{21} = \beta_{32} = \beta_{54} = \beta_{65} = 1$, $\theta_{11} = \theta_{22} = \theta_{33}$, $\theta_{44} = \theta_{55} = \theta_{66}$, the four Ψ's unconstrained. The implication of this model is that the error variances remain unchanged while the true scores and the variances of the true scores vary.
6. $\beta_{21} = \beta_{32} = \beta_{54} = \beta_{65} = 1$, the θ's unconstrained, the four Ψ's unconstrained. Not only are the error variances allowed to vary, the variances of the disturbance terms may also change. For this reason this model cannot be identified.
7. The four β's unconstrained, $\theta_{11} = \theta_{22} = \theta_{33}$, $\theta_{44} = \theta_{55} = \theta_{66}$, the four Ψ's unconstrained. The implication of this model is the same as for Model 5.
8. The four β's, and the θ's, and the four Ψ's unconstrained. This is the most general model formulated by Equations 2 and 3. This model cannot be identified.

Application of the Models

There are no a priori reasons for choosing one of the models over the others. Therefore, all the models were tested against the data. The results of the analysis are shown in Table 3. It should be noted that although Models 6 and 8 are unidentified, their goodness-of-fit statistic is still useful.

Table 3 shows that Model 1, the most restricted model, must be rejected. Models 2, 3, and 4 are acceptable at the .05 level. It should be noted that Models 2, 3, and 4 are hierarchically related, since Models 2 and 3 can be obtained from Model 4 by alternate restrictions. If one model is a restricted case of another model, the more restricted model can be tested against the less restricted one, since the difference between the two goodness-of-fit statistics is also chi-square distributed with the number of degrees of freedom equal to the difference in numbers of restrictions in the two models. However, this test holds only when the fit of the more general model is statistically acceptable. Otherwise, the difference in the goodness-of-fit statistics is not centrally chi-square distributed, as recently found by Satorra and Saris (1983).

Because Model 4 cannot be rejected on statistical grounds, it is possible to test the effects of the additional sets of restrictions of Model 2 and 3 by comparing the difference of the goodness-of-fit statistics of Models 2 and 4 with those of Models 3 and 4, respectively. These quantities of 7.14 and 6.43, respectively, are chi-square distributed with four degrees of freedom ($p > .05$ in both cases). Unfortunately, these tests do not lead to a clearer picture. Both sets of restrictions seem to be acceptable if they are introduced separately. Consequently, this analysis leads only to the conclusion that the combination of the two sets of restrictions (Model 1) is unacceptable, while the choice of Model 2 or Model 3 is arbitrary. For the moment Model 4 is to be preferred when there is a need to avoid an arbitrary decision.

The third set of restrictions, the constraints on the disturbance variances (Ψ), seems acceptable, since the comparison of Models 2, 3, and 4 with their counterparts without these constraints (Models 6, 7, and 8, respectively) shows in each case that the difference in the goodness-of-fit statistic is not significant at the .05 level. This result indicates that the constraints on the variances of the disturbance terms cannot be rejected. Given these results, Model 4 still seems the most acceptable model.

However, inspection of the parameter estimates obtained under Model 4 shows a systematic pattern

Table 3
Goodness of Fit Statistics of the Eight Different Test Equivalence Models

Model	Ψ	β	θ	Fit χ^2	Fit df	Prob. Level
1	0*	1*	constrained	30.68	16	.015
2	0*	1*	free	19.46	12	.078
3	0*	free	constrained	18.75	12	.095
4	0*	free	free	12.32	8	.137
5	free	1*	constrained	22.57	12	.032
6	free	1*	free	17.44	8	.026
7	free	free	constrained	14.32	8	.074
8	free	free	free	11.92	4	.018

*These parameter values were specified by hypothesis.

which suggests an alternative model formulated by Heise (1969). Heise has suggested that although the variances of the true and the error scores may vary over time, the ratio between them can remain unchanged. As a consequence, the reliabilities remain stable while the observed score variances change. As the results of the analysis pointed in this direction, this hypothesis was tested. According to Heise's model, all variables are standardized. This leads to

$$x_i^* = \lambda_i^* \tau_i^* + \varepsilon_i^* \quad \text{for } i = 1 \text{ to } 6, \qquad [4]$$

where

$$x_i^* = x_i/\sigma_{x_i},$$
$$\lambda_i^* = \sigma_{\tau_i}/\sigma_{x_i},$$
$$\varepsilon_i^* = \varepsilon_i/\sigma_{x_i},$$
$$\tau_i^* = \tau_i/\sigma_{\tau_i};$$

while

$$E(x_i^*) = E(\tau_i^*) = E(\varepsilon_i^*) = 0 \quad \text{for all } i,$$
$$E(\tau_i^*, \varepsilon_j^*) = 0 \quad \text{for all } i, j,$$
$$E(\varepsilon_i^*, \varepsilon_j^*) = 0 \quad \text{for all } i \neq j.$$

Since according to Model 4, $\Psi_{ii} = 0$, it follows that after standardization, the true scores have the following form

$$\tau_1^* = \tau_2^* = \tau_3^* \text{ and } \tau_4^* = \tau_5^* = \tau_6^*. \qquad [5]$$

The correlation between the two sets of true scores is denoted by $\sigma^*_{\tau_4\tau_1}$. No restriction will be made on this correlation. This means that this parameter will have to be estimated. Based on Heise (1969) it was expected that the ratio between the true score variances and the error variances would remain unchanged for the two test forms, meaning that

$$\lambda_1^* = \lambda_2^* = \lambda_3^* \text{ and } \lambda_4^* = \lambda_5^* = \lambda_6^*. \qquad [6]$$

Applying the model of Equations 4, 5, and 6 to the correlation matrix yields a chi-square of 13.26 with 16 degrees of freedom. This is exactly the same fit obtained for the model of Werts et al. (1980). This model also reproduces the expected correlation matrix among the variables as well as the model of Werts et al. Because the fit of the model is acceptable, and as will be seen, the parameter estimates are easily interpretable, it was decided to use the present authors' simplified model in preference to that of Werts et al. with its two questionable assumptions. The standardized maximum likelihood estimates of this model are shown in Table 4.

It can be seen that the correlation between the true scores of the TSWE and the essay test is not unity. The estimated correlation between the true scores is .888, and the standard error is .028. This confirms the present authors' criticism that the two tests measure different abilities. By leaving this correlation free for estimation, correlated measurement errors for the essay test are not necessary

Table 4
Standardized Maximum Likelihood Estimates of the Parameters of the "Stable Variance Ratio Model"

Parameters	Estimates	Standard Errors
$\lambda_1^* = \lambda_2^* = \lambda_3^*$	.919	.045
$\lambda_4^* \quad \lambda_5^* \quad \lambda_6^*$	.744	.044
$\sigma^*_{\tau_1\tau_4}$	.888	.028
Goodness-of-fit statistic	$\chi^2_{16} = 13.26$	$p = .653$

to obtain an acceptable fit. This means that the test-retest estimates for the reliability of the essay test are not biased. The reliability of the essay test is estimated as the squared factor loading, i.e., $(.744)^2 = .553$. This is significantly higher than reported by Werts et al. (1980), i.e., $(.676)^2 = .457$. The reliability of the TSWE measures remains the same as estimated by Werts et al., i.e., $(.919)^2 = .844$.

Discussion and Conclusions

The results of the analyses illustrate that the fit of a model cannot be the only criterion for deciding whether the best model has been found. In this study three different models were found which fitted the data equally well. In such instances the preference of one model to others depends on the plausibility of the assumptions of the models, as long as the assumptions cannot otherwise be tested. The present analyses showed that the assumption that different tests measure the same ability can and should be tested. A more elaborate discussion of this point is provided by Saris (1980).

In a longitudinal design, the assumption that the same true score underlies different tests can lead to misinterpretations, as in the case of the paper by Werts et al. (1980). The present authors show that Werts et al. needed correlated errors to correct for their implausible assumption that the TSWE and the essay test represented the same true score. They entered correlated measurement errors for the essay test, but as was pointed out, entering correlated errors for the TSWE will also lead to an acceptable fit of the model, at least in terms of the chi-square statistic and the reproduced correlation matrix. A more general model was therefore formulated to test the effects of the different assumptions. This has lead to the final model, in which two sets of perfectly correlated standardized true scores for the two test forms and stable reliabilities are postulated. Although the model of Werts et al. and this final model fit the data equally well, the final model (and interpretation) is preferable because of its less doubtful assumptions. The more restrictive hypotheses of Werts et al. thus remained open for testing. It could then be shown that their first two assumptions had to be rejected.

This exercise in model specification and model testing is not of minor importance: In the end a different estimate of the reliability of one of the two instruments used was obtained. Such a difference might influence all further results obtained from analyses performed with these instruments. Correlated errors in many studies have been introduced to correct for the unwarranted assumption of equal true scores of different tests. The last assumption, if it is incorrect, means that the error terms in the models will contain random errors as well as unique aspects of a test. Consequently, the reliability will be underestimated, and in a longitudinal study these unique aspects will lead to correlations between the measures at different points in time.

References

Breland, H. M., & Gaynor, J. L. A comparison of direct and indirect assessments of writing skill. *Journal of Educational Measurement,* 1979, *16,* 119–128.

Heise, D. R. Separating reliability and stability in test-retest correlation. *American Sociological Review,* 1969, *34,* 93–101.

Saris, W. E. Different questions, different variables, In C. P. Middendorp, B. Niemöller, & W. E. Saris (Eds.), *Sociometric Research 1980.* Amsterdam: Dutch Sociometric Society, 1980.

Satorra, A., & Saris, W. E. The accuracy of a procedure for calculating the power of the likelihood ratio test as used within the LISREL framework. In C. P Middendorp, B. Niemöller, & W. E. Saris (Eds.), *Sociometric Research 1982.* Amsterdam: Dutch Sociometric Society, 1982.

Werts, C. E., Breland, H. M., Grandy, J., & Rock, D. R. Using longitudinal data to estimate reliability in the presence of correlated measurement errors. *Educational and Psychological Measurement,* 1980, *40,* 19–29.

Wiley, D. E., & Wiley, J. A. The estimation of measurement error in panel data. *American Sociological Review,* 1970, *35,* 112–117.

Author's Address

Send requests for reprints or further information to Henk Blok, Stichting Centrum voor Onderwijsonderzoek van de Universiteit van Amsterdam, Singel 138, 1015 AG Amsterdam, The Netherlands.