Estimating Measurement Error on Highly Speeded Tests

Susan E. Whitely
University of Kansas

Despite many advances in test theory, estimating measurement error which arises from temporary changes in the person or test situation has remained relatively unchanged. Unfortunately, not only are short-term instabilities the most important source of measurement error for many traits, especially those measured from highly speeded tests; but the classical test-retest formula for estimating error is based on untenable assumptions with respect to practice effects. The current paper presents a method which gives maximum likelihood estimates of measurement error within the context of a simplex model for practice effects. The appropriateness of the model is tested for five traits, and error estimates are compared to the classical formula estimates.

Procedures for estimating measurement error in test scores due to the internal structure of the test have changed substantially since the introduction of strong true score models in psychometrics (i.e., Lord & Novick, 1968). Classical indices of measurement error, such as those from the Kuder-Richardson formulas, have been replaced by estimation errors associated with each ability parameter. Under strong true score models, such as the Birnbaum (1968) two-parameter logistic model, measurement error for internal structure is estimaied within a well-specified model, using efficient estimation methods.

However, the estimation of another important type of measurement error—short-term instabilities which arise from temporary changes in the person or testing situation—has been relatively ignored in contemporary developments. This is unfortunate for two reasons. First, for many tests, temporal fluctuations are the only meaningful source of measurement error. For example, the perceptual and motor traits which are included on multiple aptitude batteries are measured by highly speeded tests in which the items are nearly repetitions of each other. Although the internal structure of these tests is not an important source of measurement error, temporary conditions of the person or testing situation are thought to substantially influence scores. Second, the most widely used procedure for estimating measurement error due to short-term instabilities is based on untenable assumptions. Using a test-retest design, measurement error, σ_e, is "estimated" by the following formula:

$$\hat{\sigma}^2_e \equiv \hat{\sigma}^2_y (1 - r_{12}) \qquad [1]$$

where $\hat{\sigma}^2_y$ = a "pooled" estimate of observed score variance from test and retest

r_{12} = correlation between test and retest

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 3, No. 2 Spring 1979 pp. 141-154
© Copyright 1978 West Publishing Co.

As Lord and Novick (1968) point out, not only does this popular formula define sample statistics as population estimates, but it must be assumed that practice, fatigue, and memory do not influence either true or error scores on repeated testing. Although fatigue and memory effects can be controlled through experimental design of the reliability study, substantial practice effects are usually found on these highly speeded tests.

If an estimate of measurement error for short-term fluctuations is desired for the first testing, practice can potentially bias both $\hat{\sigma}_y^2$ and r_{12} in Equation 1. The pooled variance, $\hat{\sigma}_y^2$, will be inappropriate for Test 1 if variability either increases or decreases with practice. Furthermore, the correlation, r_{12}, will reflect not only the covariance of true scores, but also any covariance of Test 1 true score with change. Given the following models for observed scores on Test 1 and Test 2:

$$X_1 = \gamma + e_1 \qquad [2]$$

$$X_2 = \gamma + \alpha + e_2 \qquad [3]$$

where γ = true score
e_1 = error score for test 1
e_2 = error score for test 2
α = practice effect

and the usual assumptions about uncorrelated errors with a mean of zero, as follows:

$$\bar{e}_1 = \bar{e}_2 = 0 \qquad [4]$$

$$\sigma_{e_1e_2} = \sigma_{\gamma e_1} = \sigma_{\alpha e_1} = \sigma_{\alpha e_2} = 0 \qquad [5]$$

it can be shown that the covariance between Test 1 and Test 2 reflects both the variance of true scores and the covariance of true score and practice. Substituting for X_1 and X_2 according to Equations 2 and 3,

$$\sigma_{X_1X_2} = 1/N\sum(\gamma + e_1 + \bar{e})(\gamma + \alpha + e_2 - \bar{\gamma} - \bar{\alpha}) \qquad [6]$$

and multiplying, separating summations, and applying the assumptions given in Equations 4 and 5, the covariance of the observed scores is given as the following:

$$\sigma_{X_1X_2} = \sigma_\gamma^2 + \sigma_{\alpha\gamma} \qquad [7]$$

The two-trial estimate of true score variance, then, is completely confounded with the covariance of change and true score. Because of this confounding, a two-trial estimate of true score variance may contribute in Equation 1 to either substantially overestimating or underestimating measurement error.

Unfortunately, however, in spite of the many difficulties with the classical estimate of short-term instability, its popularity continues. A better procedure than applying the classical formula inappropriately is to efficiently estimate measurement error within a model which includes practice effects. The purpose of this paper is to present a method for estimating measurement error due to temporal fluctuations within a simplex model which requires more than two test repetitions so that the practice effect may be identified and separated from Test 1 true scores and measurement errors. The method not only uses efficient maximum likelihood estimation procedures but also tests alternative measurement models for appropriateness to the repeated testings. The method is demonstrated on multiple repeated measurements taken on five perceptual-motor traits and is compared to error estimates based on two test repetitions.

The Covariance Structure Model

The method for estimating measurement error for highly speeded tests to be developed here is based on Jöreskog's (1974) generalized analysis of covariance structure model. Basically, Jöreskog's model attempts to reproduce a covariance matrix by a structural equation, as follows:

$$\Sigma = \Lambda\phi\Lambda' + \psi^2 \qquad [8]$$

where Σ = covariance matrix
Λ, ϕ, Ψ = matrices specified by a model

In the general covariance model, the order of Λ,Φ, and Ψ^2 is smaller or equal to the order of Σ. Three types of parameters may be contained in Λ,Φ, and Ψ^2—free, fixed, or constrained parameters. The fixed and constrained parameters are the specification of the model. Jöreskog (1974) has developed maximum likelihood estimation procedures for the free and constrained parameters and a log likelihood ratio chi-square test for the appropriateness of the specified model for the data.

A special adaptation of the covariance structure model for testing fit of alternative test forms to different test theory models was given by Jöreskog (1974). True and error variances may be estimated within the context of each model.

Figure 1 gives the structural forms and number of free parameters for three test theory models—parallel, tau–equivalent, and congeneric measurements. For parallel measurements, the covariance matrix Σ is reproduced by multiplying unit vectors by a single value for Λ, λ^2, and multiplying an identity matrix by a single value for Ψ^2. The single value for Λ and Ψ results from assuming both that all true score variances (λ^2) and all error variances (ψ^2) are equal among the *t* tests. Thus, the model contains only two free parameters.

For tau-quivalent measurements, true score variances are assumed equal between the *t* tests, while error variances are unconstrained. True score variances, then, are specified as in the parallel measurement model; but error variances are estimated for each test yielding *t* + 1 parameters to be estimated. For congeneric tests, neither true nor error variances are constrained to be equal among the tests. However, the model

Figure 1

Structural Form for True Score Variances and Covariances among *t* Repeated Tests in a Quasi-Weiner Simplex

$$\Lambda\phi\Lambda' = \left[\begin{array}{llll} \phi_1^2 & & & \\ \phi_1^2 & \phi_1^2 + {}_2^2 & & \\ \phi_1^2 & \phi_1^2 + \phi_2^2 & \phi_1^2 + \phi_2^2 + \phi_3^2 & \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \\ \phi_1^2 & \phi_1^2 + \phi_2^2 & \phi_1^2 + \phi_2^2 + \phi_3^2 & \phi_1^2 + \phi_2^2 + \phi_3^2 + .. {}_t^2 \end{array}\right.$$

where ϕ_t^2 = added true score variance at repetition t does specify that the tests measure one common factor, so that t parameters are estimated for true score variances and other t parameters are estimated for error score variances. Although Jöreskog's (1974) paper links test theory to the covariance model approach, none of these traditional test theory models are formally appropriate for repeated testings which show practice effects because systematic score changes are not specified among the t tests. Even under congeneric tests, the least restrictive model, the unequal true score variances are assumed to arise for differing metrics for the t tests, rather than systematic change as a result of practice.

A structural model for covariances which does specify true change over practice is the quasi-Weiner simplex (Guttman, 1954; Jöreskog, 1974). Although this model has never before been used as a test theory model, it specifies both true and error variances, with systematic true variance change over trials of practice. Simplex models have often characterized performance on perceptual and motor tasks (Jones, 1970). Guttman (1954) presents several characteristics of correlation matrices which fit a simplex, including increasing correlations over trials and decreasing correlations away from the main diagonal.

The structural form for the true score variances and covariances among t repeated tests in a quasi-Weiner simplex is given in Figure 1. Thus, true score variance increases over trials in this model, due to added components of performance. In Jöreskog's general structural model, given in Equation 8, the quasi-Weiner simplex is specified when Φ is a diagonal matrix of the additive true score variances Φ_t^2; Ψ^2 is a diagonal matrix of the error variances for each repetition; and Λ is a unit lower triangular matrix as follows:

$$\Lambda = \begin{bmatrix} 1 & & & \\ 1 & 1 & & \\ 1 & 1 & 1 & \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \\ 1 & 1 & 1 \cdots & 1 \end{bmatrix}$$

Presented in Table 1 are three alternative structural models for a Weiner simplex which vary in assumptions about measurement error. The first model, the pure Weiner simplex, specifies no measurement error so that only the t additive true score variances are estimated. The second model specifies measurement error for each repetition; but only $t + 1$ parameters are estimated, since error variance is constrained to be equal for each trial. The third model is similar to the second except that error variances are not assumed to be equal. Logically, this model should specify a separate true score variance and error variance for each trial. However, only $2t$-1 parameters can be identified for this model, as the last trial true score variance and error variance cannot be separated. Other possible models for a simplex would result from permitting some correlated errors in the Ψ^2 matrix, such as non-zero error correlations between adjacent trials.

The quasi-Weiner simplex provides a plausible model for estimating error variance due to short-term instabilities for several reasons. First, the Weiner simplex specifies systemmatic variance increases over repeated testings, as is commonly observed in test repetitions. Second, true and error variances may be separately estimated for each repetition. Third, possible overestimation of true score variances due to correlated errors among temporally close repetitions may be corrected by specifying correlated error models; correlated errors could arise from emotional or physical conditions of the person or situational characteristics of the test environment that persist over at least two testings. Fourth, the multivariate normality assumption of Jöreskog's (1974) covariance model seems generally reasonable for carefully constructed test measures.

Successfully applying a quasi-Weiner simplex model to estimate measurement error over time requires controlling for some possible confounding influences through experimental design. First, the repeated testings must be spaced so that fatigue does not influence per-

Table 1
Schematic for Test Models
$\Sigma = \Lambda\phi\Lambda' + \psi^2$

Test Model	Structural Model	Number of Parameters
Parallel measurements	$\Sigma = \lambda^2 11' + \psi^2 I$	2
Tau equivalent measurements	$\Sigma = \lambda^2 \overline{11}' + \psi^2$	t + 1
Congeneric measurements	$\Sigma = \lambda\lambda' + \psi^2$	2t
Simplex model		
Weiner simplex	$\Sigma = \Lambda\phi\Lambda'$	t
Quasi-Weiner simplex, equal ψ^2	$\Sigma = \Lambda\phi\Lambda' + \Psi^2 I$	t + 1
Quasi-Weiner simplex, free ψ^2	$\Sigma = \Lambda\phi\Lambda' + \Psi^2$	2t - 1

formance. Second, the influence of memory for or adaptation to specific "items" must be carefully considered. The trait measured by t^{th} repetition of the same test may be quite different than the trait measured by the t^{th} testing from t parallel forms. Performance on the t^{th} repetition for the former may well depend heavily on rote skills, thus representing few aspects of the initial test performance. The Weiner simplex models would seem to be most appropriate for repetitions from parallel forms, as the abilities required in the initial testing are postulated to continue in later testings.

The following sections examine the relative appropriateness of the Weiner simplex models and the test theory models for estimating measurement error due to short-term instabilities on highly speeded tests.

Method

The test for model appropriateness in estimating short-term instabilities is given by a reanalysis of Berdie's (1969) data on repeated measurements. Although these data were originally collected to examine intraindividual variability as a measurement construct, the carefully controlled experimental design makes these data very appropriate for the purposes of the current study. Since the methods are described in Berdie (1969), only a brief summary with respect to the current problem will be given here.

Tests

The tests used in the Berdie (1969) study were 20 alternative forms for each of 6 perceptual and motor traits. Standardization data were given for 5 of these traits by Moran, Kimble, and Mefferd (1964).

The five traits for which standardization data are available include three of Thurstone's perception factors—Flexibility of Closure, Speed of Perception, and Speed of Closure—and two traits from the French (1966) Kit of Reference Factors—Number Facility and Visualization. The tests are all highly speeded and scored as number correct within a short time interval of 2-1/2 to 3 minutes.

Moran, Kimble, and Mefferd (1964) found small, but significant, mean differences between forms. Scores on the forms were adjusted for the mean differences to provide an unbiased estimate of generic true scores (as defined by Lord & Novick, 1968) from each form. Although possible interactions of generic true scores and error remained, it was assumed for convenience that the adjusted scores are from rigorously parallel forms.

The subjects were 79 undergraduates enrolled in the Institute of Technology at the University of Minnesota. These subjects were paid to participate in the experiment and were screened from a larger pool of applicants on the basis of availability for the 20 scheduled testing sessions.

Procedure

The tests were administered on 20 consecutive days to 79 undergraduates. The 24-hour interval eliminated fatigue as a factor contributing to instability, and the administration of a different form on each occasion controlled for memory effects. Forms were counterbalanced over days to control for any uncorrected differences between the forms.

Design for Reliability

The time intervals of 20 daily testings controlled for fatigue as a factor in performance instability. The use of 20 alternative forms, rather than 20 repetitions of a single test, controlled for memory or adaptation effects. The remaining contributions to measurement error which could be estimated, then, were the specific testing conditions and short-term fluctuations within the persons tested.

Results

The nature of the practice effect was determined by examining the daily means and variances for each trait. Figure 2 shows sharp mean increases for all five traits. Similarly, Figures 3 and 4 show sharp increases for the variances for all but one trait. For Speed of Perception, variances were gradually decreasing. Thus, practice was found to substantially influence observed scores on these five traits.

Table 2 presents the log likelihood ratio χ^2's obtained from fitting some test theory models to the covariances between repetitions for each of five tests. It can be seen for each test that better fit was obtained for the less restrictive test models (especially congeneric tests). The difference in χ^2 between the parallel and tau-equivalent test models were 49.59, 25.11, 93.27, 69.79, and 53.21 in the trait order given in Table 1. Using the χ^2 difference test given by Jöreskog (1974), all but the second value were significant at $p <$ .05, with 19 degrees of freedom. The χ^2 differences between the tau-equivalent and congeneric models were 81.87, 30.35, 39.01, 158.97, and 122.33, respectively, for the trait order given in Table 1. With 19 degrees of freedom, all of these differences were statistically significant at $p < .05$.

It should be noted that not even the least restrictive model, congeneric tests, fit any trait very well. The sample covariances differed significantly (p's $< .05$) from the covariances estimated by the model, and all the χ^2's were quite large. However, as indicated by Jöreskog (1974), a model may still be useful even if the data deviate somewhat from prediction. Another way to examine the data is by the ratio of chi-square to degrees of freedom. For the congeneric model in Table 2, these ratios were 12.25, 7.66, 20.58, 12.56, and 5.51, respectively. These ratios, with the possible exception of the Flexibility of Closure Test, were too large for the congeneric model to be useful.

Table 3 presents the χ^2 obtained from fitting four simplex models to the 20 repetitions of each trait. The perfect simplex model, with no error, fit quite poorly for all traits, as indicated by the large χ^2 values. The equal error model, which estimates only one more parameter than the perfect simplex model, led to substantially smaller χ^2's. The χ^2 differences were 548.45, 590.29, 906.53, 543.15, and 595.01. With 1 *df*, these differences were highly significant (p's $< .0001$).

Allowing error variances to be unequal resulted in the χ^2's in the third model given in Table 3. The χ^2 differences from the equal error model were 35.46, 29.22, 95.99, 68.85, and 53.16, respectively. With 18 *df*, all these differences were significant at $p < .05$. The last model given in Table 3 permits correlated errors between adjacent trials in addition to unequal error variances. The χ^2 differences of this model from the unequal error model were 42.06, 24.39, 122.03, 33.88, and 38.39, respectively. With 18 *df*, all but the second trait were significant at $p <$ .05.

Figure 2
Plot of Daily Means for Five Traits

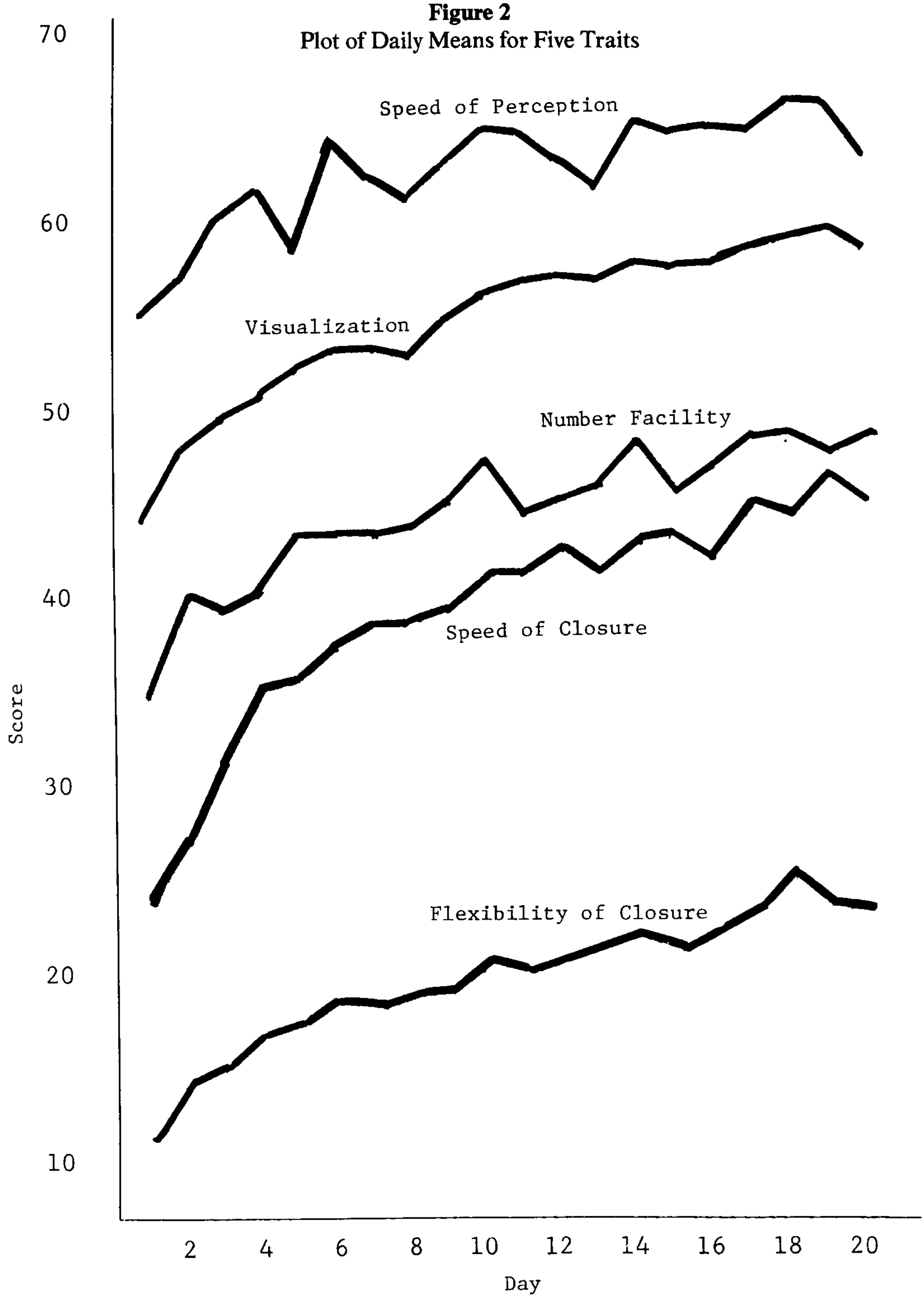

Figure 3
Plot of Daily Variances for Two Traits

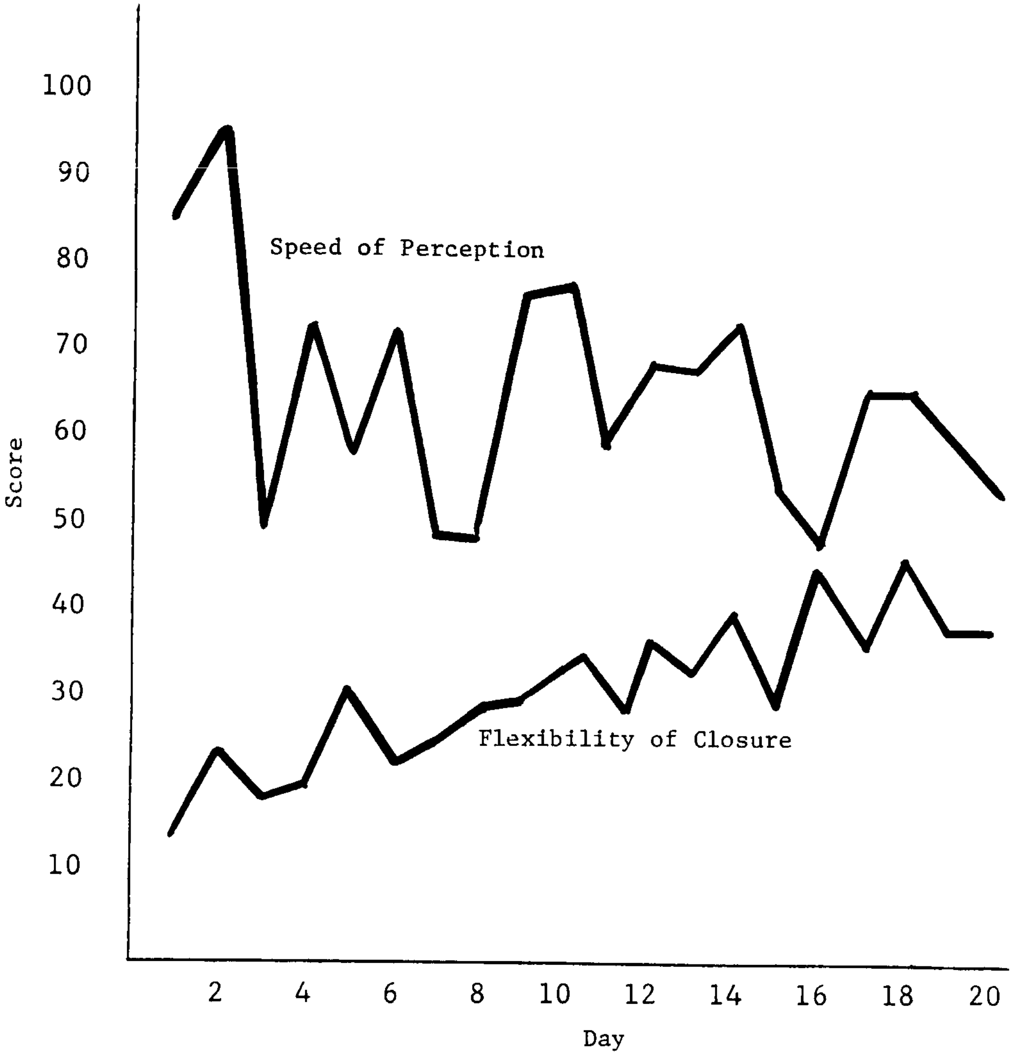

Figure 4
Plot of Daily Variances for Three Traits

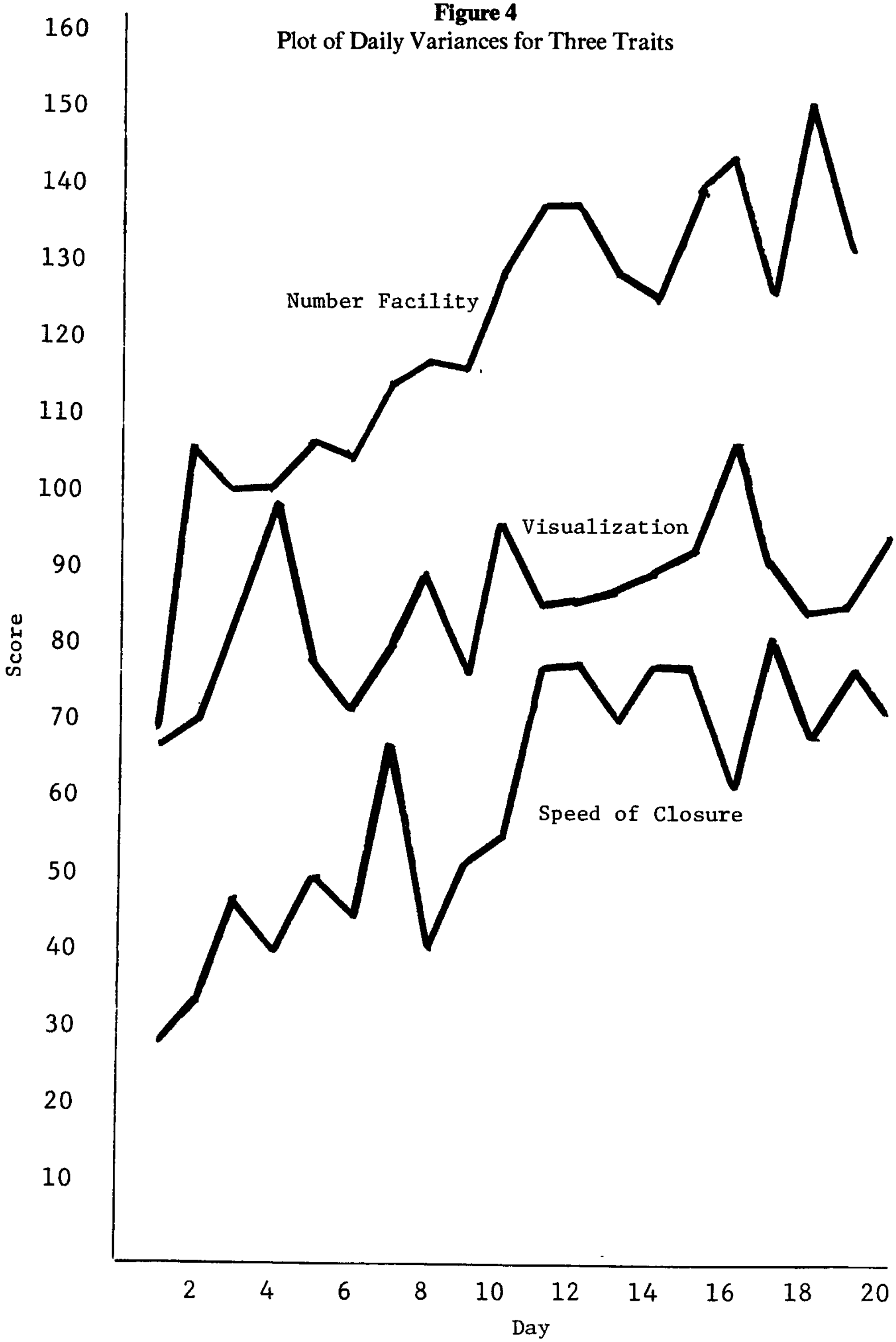

Table 2
Summary Analysis for Test Theory Covariance Models

	Parallel (p=2; df=208)		Tau-Equivalent (p=21; df=189)		Congeneric (p=40; df=170)	
Trait	X^2	p	X^2	p	X^2	p
Number Facility	621.83	.0000	572.24	.0000	490.37	.0000
Visualization	362.05	.0000	336.94	.0000	306.59	.0000
Speed of Perception	955.60	.0000	862.33	.0000	823.32	.0000
Speed of Closure	731.24	.0000	661.45	.0000	502.48	.0000
Flexibility of Closure	396.03	.0000	342.82	.0000	220.49	.0000

If $p < .05$ is the criterion for model fit, only Flexibility of Closure was not significantly different from the simplex model with correlated error. If $p < .01$ is the criterion, fit was also achieved for Visualization. To examine model usefulness, the ratios of χ^2 to *df* were computed. These values were 3.81, 3.32, 11.46, 6.28, and 3.09, respectively. The χ^2 ratio of 3.81 for Number Facility was quite close to the X^2 ratios for Flexibility of Closure and Visualization, indicating that the model is probably useful for this trait as well. However, the large values obtained for Speed of Perception indicate that an appropriate model has not been specified. The Speed of Closure χ^2 ratio was between these extremes, indicating marginal usefulness.

Comparing the simplex models to the test theory models is appropriate for those models with approximately the same number of parameters. Thus, the tau-equivalent model may be compared to the simplex model with equal error, at 21 *df* each; and the congeneric model may be compared to the simplex model with unequal error, at 40 and 39 *df*, respectively. Although an appropriate significance test is not available to compare across models, an inspection of the χ^2's revealed that substantially better fit was obtained for all five traits by the simplex model with equal error and for four of the five traits by the simplex model with unequal errors. The exception for the latter model was Flexibility of Closure; for while the χ^2 for the simplex was smaller, the difference between the χ^2 values was not very large.

Table 4 presents estimates of error variance obtained from the classical stability estimate and the alternative covariance models for the five traits. It can be seen that the simplex models estimated smaller error variances when compared to test theory models with approximately equal *df*. Also of interest in Table 4 is that the best-fitting simplex models—the unequal error model and the correlated model—gave uniformly smaller error variance estimates than the classical two-trial formula. The traditional test models, on the other hand, gave generally larger error variance estimates than the classical formula.

Although the simplex models led to better fits and smaller error variances than test theory models, the final values of χ^2 were still too large to clearly indicate the utility of the procedure. It is possible that 20 repetitions are too many. Fewer repetitions may provide better estimates because motivational factors, such as boredom, may be confounding the covariances when performance is continued to 20 trials. If the possibility of correlated error is anticipated, five repetitions give sufficient degrees of freedom to test the simplex model and to estimate error variances. Table 5 shows the tests of fit and error estimates from the simplex model with unequal errors for the first five trials. With the exception of Speed of Perception—the test with decreasing variances—fit was achieved by the simplex model, showing the utility of the model for the first five trials. Inspection of the trial and error estimates, which are also presented in Table 5, showed very similar results to the 20 trial estimates.

Discussion

The problem of estimating measurement error due to short-term fluctuations was examined for five highly speeded tests of perceptual and motor traits. On all 5 traits, substantial changes in both means and variances were observed over the 20 daily testings. The classical test-retest formula for estimating measurement error is not formally appropriate for these data because, as shown in the introduction, the classical formula can be expected to give substantially biased estimates when practice influences scores.

Several alternative structural models for the test repetitions were examined for fit. The data covariances differed significantly from all the structural models from test theory. The lack of fit for parallel measurements could easily be expected from changing variances over trials, as the model constrains all true and error variances to be equal. However, since neither of the less restrictive models—the tau-equivalent test and the congeneric tests—yielded good fit, apparently practice effects scores in ways which cannot be accounted for by these models.

The simplex models yielded better fit of the data. Over 20 repetitions, significantly better fit was achieved for models which specified unequal error variances and correlations among errors for adjacent repetitions. An examination of the *p* values and chi-squares for the best simplex model with correlated errors showed fit

Table 3

Summary Analysis for Simplex Structure Covariance Models

Trait	Perfect Simplex (p=20; df=190)		Equal Errors (p=21; df=189)		Unequal Errors (p=39; df=171)		Correlated Errors (p=57; df=153)	
	χ^2	p	χ^2	p	χ^2	p	χ^2	p
Number Facility	842.96	.0000	294.51	.0000	259.05	.0000	216.99	.0005
Visualization	833.18	.0000	242.89	.0050	213.67	.0148	189.28	.0245
Speed of Perception	1778.21	.0000	871.68	.0000	775.69	.0000	653.66	.0000
Speed of Closure	1004.28	.0000	461.13	.0000	392.28	.0000	358.40	.0000
Flexibility of Closure	862.85	.0000	267.84	.0000	214.68	.0132	176.29	.0956

Table 4

Estimates of First Test Measurement Error from Alternative Test Models

Model	Number of Parameters	Number Facility	Visualization	Speed of Perception	Speed of Closure	Flexibility of Closure
Classical Two-Trial "Estimate"	--	11.92	15.45	47.82	9.10	7.20
$\sigma^2_{e_1} \equiv s^2_y(1-r_{12})$						
Estimates from 20 Replications						
$(\sigma^2_{e_1} = \psi^2_{11})$						
Traditional test models						
Parallel measurements	2	16.92	20.34	28.31	19.13	7.66
Tau equivalent measurements	21	31.70	26.28	57.49	34.82	10.11
Congeneric measurements	40	20.69	24.08	57.48	20.21	7.02
Simplex models						
Weiner simplex, equal error	21	10.59	16.71	26.41	12.57	6.28
Weiner simplex, uncorrelated error	39	3.91	11.32	46.31	5.58	3.39
Weiner simplex, correlated error	57	3.32	11.03	43.60	6.44	3.55

for two traits—Visualization and Flexibility of Closure—and probable usefulness for Number Facility. The appropriateness of the simplex for Speed of Closure was marginal, while the model clearly was not appropriate for Speed of Perception.

When compared to test theory models with approximately equal numbers of parameters, the simplex models always fit better, and usually substantially better. Although both the less restrictive test theory models and the simplex models can account for unequal variances and covariances among the repetitions, only the simplex models specify systematic changes as a result of practice. The quasi-Weiner simplexes which were fit specify additive components of individual differences over practice which persist over the following trials. Thus, the simplex models specify true change over practice, rather than merely changing error variances or changing measurement scales for true scores. The current study strongly indicates that the simplex model gives a better account of individual differences over repeated testings.

Although the simplex models provided better accounts of the data than the test theory models, the general usefulness of the model was questionable, since fit was achieved only for two tests. Since 20 repetitions of the same measures may eventually introduce confounding motivational factors, such as boredom, fit of the simplex model was examined for the first 5 repetitions only. Fit was clearly attained for the simplex model with unequal error variances for all traits except Speed Perception.

Thus, the five-repetitions data support the general applicability of the simplex model for separating true changes from error variance over repeated measurement. The failure to achieve fit for Speed of Perception indicates that some other kind of model may be appropriate for these data. This trait, unlike the others, showed *decreasing* variances over trials. The quasi-Weiner simplex model used in the current study specifies *increasing* variances.

Comparing the estimates of measurement error given by the simplex models to the classical two-trial estimates showed that error from the latter is substantially overestimated. For all traits, the simplex models led to smaller error variances. This occurs because the classical estimate attributes the practice effect observed on repeated testing to measurement error. Similar to the classical estimate, the traditional test models also lead to overestimation of error because measurement error is confounded with the practice effect.

The current study demonstrates the inappropriateness of the classical test-retest estimate of error variance for tests with substantial practice effects. It also shows that measurement error due to short-term fluctuations may be more appropriately estimated within a model for practice effects—the quasi-Weiner simplex.

Table 5
Simplex Structure Tests and Error
Variances from Unequal Error Model for the First Five Traits

	χ^2	df	p	Trial 1 Error Variance
Number Facility (3)	3.15	6	.79	4.04
Visualization (6)	4.21	6	.65	11.02
Speed of Perception (4)	24.63	6	.00	46.50
Speed of Closure (5)	9.14	6	.17	5.50
Flexibility of Closure (2)	12.46	6	.06	3.57

In actual applications of the simplex covariance model, the current results indicate that 5 test repetitions provide more satisfactory estimates of error than 20 repetitions. Five repetitions are not only more feasible in most applications, but five replications give sufficient degrees of freedom to test the data for fit to the model and estimate the parameters, without the introduction of boredom effects in performance.

References

Berdie, R. F. Intra-individual temporal variability and predictability. *Educational and Psychological Measurement*, 1969, *29*, 235–257.

Birnbaum, A. Some latent trait models. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 1968.

French, J. W. *Kit of reference tests for cognitive factors (rev. ed.)*. Princeton, N.J.: Educational Testing Service, 1966.

Guttman, L. A new approach to factor analysis: The radex. In P. E. Lazarfeld (Ed.), *Mathematical thinking in the social sciences*. Glencoe, IL: Free Press, 1954.

Jones, M. B. A two-process theory of individual differences in motor learning. *Psychological Review*, 1970, *77*, 353–360.

Jöreskog, K. G. Analyzing psychological data by structural analysis of covariance matrices. In D. H. Kranz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology*. San Francisco: Freeman, 1974.

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

Moran, L. J., Kimble, J. P., & Mefferd, R. B. Repetitive psychometric measures: Equating alternative forms. *Psychological Reports*, 1964, *14*, 335–338.

Acknowledgments

This study was partially supported by grant number 3472-x0-0038 from the University of Kansas General Research Fund, Susan E. Whitely, Principal Investigator.

Author's Address

Send requests for reprints or further information to Susan E. Whitely, Department of Psychology, University of Kansas, Lawrence, KS 66045.