Adaptive and Conventional Versions of the DAT: The First Complete Test Battery Comparison

Susan J. Henly and Kelli J. Klebe, University of Minnesota

James R. McBride, The Psychological Corporation

Robert Cudeck, University of Minnesota

A group of covariance structure models was examined to ascertain the similarity between conventionally administered and computerized adaptive (CAT) versions of the complete battery of the Differential Aptitude Tests (DAT). Two factor analysis models developed from classical test theory and three models with a multiplicative structure for these multitrait-multimethod data were developed and then fit to sample data in a double cross-validation design. All three direct-product models performed better than the factor analysis models in both calibration and cross-validation subsamples. The cross-validated, disattenuated correlation between the administration methods in the best-performing direct-product model was very high in both groups (.98 and .97), suggesting that the CAT version of the DAT is an adequate representation of the conventional test battery. However, some evidence suggested that there are substantial differences between the printed and computerized versions of the one speeded test in the battery. *Index terms: adaptive tests, computerized adaptive testing, covariance structure, cross-validation, Differential Aptitude Tests, direct-product models, factor analysis, multitrait-multimethod matrices.*

The majority of studies that have compared scores from conventional paper-and-pencil tests with scores from tailored or computerized adaptive (CAT) versions of the same tests (Lord, 1974; 1980, chap. 10) have focused on a small subset of scales from a complete battery. Indeed, the bulk of reported research on this topic has been concerned with comparisons of single conventional tests and an adaptive version designed to measure the same ability. Sympson, Weiss, and Ree (1982, pp. 1–2) briefly reviewed some recent literature on this kind of comparison.

Due to the success with which single tests have been converted into adaptive forms, the logical next step is to adapt complete test batteries. Conventional test batteries with a history of use in applied settings are an appropriate choice for translation into CAT versions.

To date, the most thoroughly studied partial battery of tests in adaptive form is the Armed Services Vocational Aptitude Battery (ASVAB; U.S. Department of Defense, 1982). Although a completely computerized version of the ASVAB has been developed, only certain subtests of the conventional and adaptive versions have been formally compared. For example, Moreno, Wetzel, McBride, and Weiss (1984) evaluated the Arithmetic Reasoning, Word Knowledge, and Paragraph Comprehension subtests. Cudeck (1985) compared conventional and adaptive versions of the Arithmetic Reasoning, Word Knowledge, General Science, and Mathematics Knowledge subtests. Both of these studies reported very favorable correspondence between the conventional and adaptive versions of the subtests.

However, the particular subtests that were eval-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 13, No. 4, December 1989, pp. 363–371

0146-6216/89/040363-09$1.70

uated in these studies are generally regarded as being among the most reliable and valid of the battery. These optimistic results may not be representative of the kind of performance that can be expected from other ASVAB subtests or from other test batteries.

Moreover, certain subtests have been excluded from comparative studies of this kind because adaptive versions present certain difficulties. For example, the ASVAB contains two speeded tests, Numerical Operations and Coding Speed, which consist of many easy items. These subtests will probably not be adapted because the items do not lend themselves to tailoring in the way that the other subtests do. Instead, the paper-and-pencil versions of the items will simply be administered by a computer in essentially a conventional manner.

The nonadaptive computerization of the Numerical Operations and Coding Speed tests would seem to present few potential difficulties. However, Greaud and Green (1986) noted that there is "no assurance" that scores based on computer presentation will be comparable to those obtained from a conventional test. They reported, for example, that ratio scores—such as the average number of correct responses per minute—were more reliable for the computer-administered test than were conventional number-correct scores. Examinees worked faster in computer mode. Even seemingly trivial changes in task with computer administration (e.g., presenting clerical coding items individually rather than in groups of seven) resulted in a low correlation between the conventional and CAT versions.

The general question that arises from these studies is whether the correspondence between a complete battery of conventional tests and an associated battery of adaptive tests will still be strong when all subtests are included. A specific issue that apparently has not yet been addressed is whether the composite structure of the battery remains the same when the adaptive version of a battery contains one or more subtests that are simply computerized replicas of their conventional test counterparts. The purpose of this study was to investigate the structural similarity between adaptive and conventional versions of the complete battery of Differential Aptitude Tests (DAT; Bennett, Seashore, & Wesman, 1982).

Structural Models of Similarity

Various approaches can be used to investigate the correspondence between subtests of two versions of a test battery (Gulliksen, 1968). The conjecture that test scores are in some way related often implies a structural model for the matrix of covariances among all the tests. Several covariance structures have been developed that are relevant for assessing the similarity between batteries of tests. Some of the most important are based on concepts from classical test theory and the study of parallel tests. Because each adaptive subtest from the DAT was designed to measure the same aptitude as the associated conventional test, it would seem obvious that the model for parallel tests would be a reasonable choice for the present purposes.

Perhaps surprisingly, however, this structure was found to be completely inappropriate in this context. The covariance structure for parallel tests (Jöreskog, 1971) specifies that each group of related tests has equal true-score variances and equal error-score variances. Unlike conventional tests, however, adaptive tests developed from item response theory models do not have a "natural" scale. Instead, variances of adaptive tests are arbitrarily fixed, frequently at unity, for some population. In the present case, the adaptive tests had scales that are functions of the associated conventional tests that were determined during equating. Therefore, because adaptive test variances are not independent functions of item responses, the classical test theory models of parallel measurements—and of related models, such as that for essentially tau-equivalent tests—are inappropriate. This conclusion appears somewhat ironic at first glance in that adaptive tests, although in this case specifically designed to be as similar as possible to conventional tests, are fundamentally unsuitable for these "strong" models of similarity.

As an alternative, consider the class of factor analysis structures for p variables written as

$$\boldsymbol{\Sigma} = \mathbf{D}_\sigma(\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi})\mathbf{D}_\sigma \quad , \tag{1}$$

subject to the additional restrictions

$$\text{diag}(\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}) = \mathbf{I} \quad , \tag{2}$$

where the matrix $\mathbf{D}_\sigma = \text{diag}(\sigma_1, \ldots, \sigma_p)$ contains scaling terms, and $\boldsymbol{\Lambda}(p \times k) = \{\lambda_{ij}\}$, $\boldsymbol{\Phi}(k \times k)$, and $\boldsymbol{\Psi} = \text{diag}(\psi_1, \ldots, \psi_p)$ are matrices of factor regression coefficients, factor covariances, and uniquenesses, respectively. This structure is scale invariant (Cudeck, 1989) and is therefore suitable for variables with possibly very different variances. In particular, it is appropriate for comparisons of conventional and adaptive tests where the variances of the latter are fixed at unity, are functions of the associated conventional tests, or are determined in some other manner. In these analyses, $\boldsymbol{\Sigma}$ also displays the form of a multitrait-multimethod (MTMM) covariance matrix, in which the traits are the subtests of the test battery and the methods correspond to conventional and adaptive versions.

Factor Analysis Models

Two special cases of the model described by Equations 1 and 2 are of interest. The first is the well-known model for congeneric tests (Jöreskog, 1971), in which the restriction $\text{diag}(\boldsymbol{\Phi}) = \mathbf{I}$ and the pattern of factor loadings

$$\lambda_{ij} = \begin{cases} \text{free for } \lambda_{lq},\ q = 1, \ldots, t; \\ \qquad\qquad l = q, q+t, \ldots, q+(m-1)t \\ 0 \text{ otherwise} \end{cases} \tag{3}$$

is imposed, where t is the number of subtests in the test battery, and m is the number of methods of test administration. This model simply specifies that the adaptive tests measure the same aptitudes as the corresponding conventional tests.

A more restrictive model imposes the t equality constraints

$$\lambda_{jj} = \lambda_{j+t,j} = \ldots = \lambda_{j+(m-1)t,j} \quad , \tag{4}$$

for $j = 1, \ldots, t$. This model is analogous to a model of tau equivalence, but because of the restriction in Equation 2 it simply assesses the extent to which the common-score variances for pairs of variables are equal. Also, because of the restriction of pairwise equality of factor loadings, Equation 2 further implies that unique variances are simultaneously equal:

$$\psi_j = \psi_{j+t} = \ldots = \psi_{j+(m-1)t} \quad , \tag{5}$$

for each of the traits $j = 1, \ldots, t$.

Direct-Product Models

Browne (1984) described a class of models for MTMM matrices that posits a multiplicative structure for the relationship between trait and method components. The least constrained structure is

$$\boldsymbol{\Sigma} = \mathbf{D}_\zeta(\mathbf{P}_m \otimes \mathbf{P}_t + \mathbf{D}_\eta^2)\mathbf{D}_\zeta \quad , \tag{6}$$

where $\mathbf{P}_m$ is of order $m \times m$ and contains correlations among method true scores, $\mathbf{P}_t$ is of order $t \times t$ and contains trait true-score correlations, and $\otimes$ indicates the Kronecker product. Elements of the diagonal matrix $\mathbf{D}_\zeta$ are scaling terms for the observed scores, while elements of the diagonal matrix $\mathbf{D}_\eta^2$ are ratios of unique-score standard deviations to common-score standard deviations. This model is referred to as the composite direct-product model with no restrictions on $\mathbf{D}_\zeta$ or $\mathbf{D}_\eta^2$ (CDPZE).

The model includes special cases with multiplicative structures for $\mathbf{D}_\zeta$ and $\mathbf{D}_\eta^2$. The most restrictive model (CDP) defines a multiplicative structure for both $\mathbf{D}_\zeta$ and $\mathbf{D}_\eta^2$:

$$\mathbf{D}_\zeta = \mathbf{D}_{\zeta(m)} \otimes \mathbf{D}_{\zeta(t)} \tag{7}$$

$$\mathbf{D}_\eta^2 = \mathbf{D}_{\eta(m)}^2 \otimes \mathbf{D}_{\eta(t)}^2 \quad , \tag{8}$$

where the diagonal matrices $\mathbf{D}_{\zeta(m)}$ and $\mathbf{D}_{\eta(m)}^2$ are of order m, and the diagonal matrices $\mathbf{D}_{\zeta(t)}$ and $\mathbf{D}_{\eta(t)}^2$ are of order t. One element in each of $\mathbf{D}_{\zeta(m)}$ and $\mathbf{D}_{\eta(m)}^2$ is fixed at unity for identification purposes. A less restrictive special case (CDPZ) defines a multiplicative structure for $\mathbf{D}_\eta^2$ only, with $\mathbf{D}_\zeta$ unconstrained.

Like the model for congeneric tests, the direct-product models estimate the true-score correlations among the abilities measured by the subtests. A useful feature of the direct-product structures that is not shared by the factor analysis models is that the former provide an overall estimate of the correlation between the two methods, providing useful

information about the degree to which the batteries are similar.

Model Selection

The purpose of fitting a model to a covariance matrix is to summarize the elements of the matrix in terms of a smaller number of parameters, thereby aiding understanding of the data (Browne, 1984). The most useful models are those with interpretable parameters that closely reproduce the observed matrix.

The classical approach to estimating model parameters and testing the probability of the implied covariance structure is now well understood (Jöreskog, 1978; Lawley & Maxwell, 1971). In practice, however, the problem of assessing the plausibility of a model is usually not straightforward or automatic. As a result, recent work has been devoted to developing and justifying various indices of fit between a sample and a reproduced covariance matrix (Akaike, 1987; Bentler & Bonett, 1980; James, Mulaik, & Brett, 1982; Tanaka & Huba, 1985). Although these indices differ from each other in significant ways, they have the common feature of attempting to identify a model that most reasonably accounts for data obtained from one sample.

As an alternative, Cudeck and Browne (1983) suggested a model selection procedure based on empirical cross-validation. The primary justification for cross-validation is that performance in future samples is a more important criterion for evaluating a model than is the ability to account for data in the sample, which is also used to estimate the model parameters.

Let $\mathbf{S}$ be the unbiased estimate of the population covariance matrix, and let Σ_k be the population covariance matrix implied by the kth model in a set of models of interest. An estimate of the population covariance matrix under the kth model, $\hat{\Sigma}_k$, is obtained using the Maximum Wishart Likelihood discrepancy function

$$F(\mathbf{S},\hat{\Sigma}) = \ln|\hat{\Sigma}| - \ln|\mathbf{S}| + \mathrm{tr}\{\mathbf{S}\hat{\Sigma}^{-1}\} - p \quad . \qquad (9)$$

In a double cross-validation study, two distinct sample covariance matrices, $\mathbf{S}_A$ and $\mathbf{S}_B$, are obtained. The model parameters are first estimated using data from both samples, computing $F(\mathbf{S}_A,\hat{\Sigma}_k)$ and $F(\mathbf{S}_B,\hat{\Sigma}_k)$ for $k = 1, \ldots, g$. Define $\hat{\Sigma}_{k|A}$ and $\hat{\Sigma}_{k|B}$ to be the estimated population covariance matrices implied by model k for Sample A and Sample B, respectively. To assess the performance of the models in another context, $F(\mathbf{S}_B,\hat{\Sigma}_{k|A})$ and $F(\mathbf{S}_A,\hat{\Sigma}_{k|B})$ are computed. The model associated with the smallest cross-validation index is considered the most effective representation from the set of structures examined. Although it is generally the case that the model with the largest number of parameters will have the smallest discrepancy function in the calibration samples, this will not necessarily hold for the validation samples.

When the number of parameters to be estimated is fairly large, as in the models studied here, there is the possibility that sample characteristics may influence the parameter estimates if sample sizes are too small or if distributions are less than optimal (Tanaka, 1987). Cross-validation circumvents this problem by assessing the performance of a model, and thus the parameter estimates, in future samples. In this way, models which are strongly influenced by chance fluctuations will not necessarily perform well in the validation samples and will be rejected in favor of models that do not capitalize on sample characteristics.

A single-sample cross-validation index for covariance structures which approximates empirical cross-validation has recently been developed (Browne & Cudeck, in press). This index can be used in situations where it is difficult to obtain two samples of reasonable size, but it is not meant as a replacement for empirical replication.

Method

The DAT

The DAT is a battery of eight ability tests designed for use in educational placement and vocational counseling in junior and senior high schools. Previous factor-analytic results of ability data were used to guide development of subtests for the DAT that represented well-recognized vocational or educational areas (Anastasi, 1988). Seven "power"

tests are included in the battery: Verbal Reasoning (VR), Numerical Ability (NA), Abstract Reasoning (AR), Mechanical Reasoning (MR), Space Relations (SR), Spelling (SP), and Language Usage (LU). The eighth test, Clerical Speed and Accuracy (CSA), is speeded.

A computerized adaptive edition of the DAT has recently been released (McBride, 1986). The seven power tests are tailored, but the speeded CSA test is merely modified for computerized administration. The eight computerized subtests used in this study were developed from the items of Form V of the DAT and implemented as described below. The initial item statistics were computed with the Rasch (1966) model using the 1982 standardization sample. The Rasch model was used because in comparative analyses the item parameters from this model yielded results that were generally as good as, or slightly better than, the results obtained from the three-parameter logistic model when evaluated in terms of both their correlations with independent ability measures (DAT Form W raw scores) and their equating accuracy (McBride, Corpe, & Wing, 1987).

In actual testing, estimates of ability at each step were calculated using a Bayesian updating technique (Owen, 1975). Items were selected by maximizing information over the items not yet encountered. Each of the adaptive tests terminated when the number of items administered was half the length of the corresponding conventional test. The adaptive tests were administered on Apple II computers. Ability estimates from the adaptive tests were reexpressed as equivalent raw scores of the conventional tests by equipercentile equating (Braun & Holland, 1982) to Form W versions (McBride et al., 1987); these equated scores were used in all analyses.

Examinees

Data for this study were gathered during the initial field test of the adaptive version of the DAT. Examinees were administered the entire DAT test battery in conventional and adaptive modes. Form W was used for the paper-and-pencil test. Order of administration was counterbalanced. Twelve school districts around the nation participated in the field test. More than 500 students, primarily in grades 8 through 12, participated. Complete scores available for 332 examinees were used for the analysis reported here (see McBride, 1986, for details).

Design

The examinees were randomly divided into two subsamples of size $n_A = 171$ and $n_B = 161$. Tables 1 and 2 list the correlation matrices, means, and standard deviations for each group. The five models evaluated are listed in Table 3, along with values of the discrepancy functions and cross-validation indices for each model in both samples. The Maximum Wishart Likelihood discrepancy function was used to estimate parameters of the models and to estimate $F(\mathbf{S}_A, \hat{\Sigma}_{k|A})$ and $F(\mathbf{S}_B, \hat{\Sigma}_{k|B})$ for each group. Cross-validation indices $F(\mathbf{S}_A, \hat{\Sigma}_{k|B})$ and $F(\mathbf{S}_B, \hat{\Sigma}_{k|A})$ were then calculated.

Results

For Sample A, all three direct-product models performed better (i.e., the discrepancy and cross-validation indices were smaller) than the factor analysis models during both calibration and cross-validation. For Sample B, the least constrained direct-product model (CDPZE) fit best during calibration, followed by the model for congeneric tests. The results on cross-validation were similar to those based on Sample A: The direct-product models consistently performed better. In this case, however, the cross-validation index was smaller for the most constrained direct-product model (CDP) than for the less restrictive basic direct-product model (CDPZE).

Although the double cross-validation procedure does indicate a set of models which should perform better in future samples, it does not imply that there is one specific model which is "the best." In order to select a single model which provides the best summary of the data, other criteria such as interpretability and parsimony need to be considered. In this case, the set of direct-product models seemed preferable to the factor analysis models with re-

Table 1
Observed DAT Correlations for Sample A (n_A = 171) and Sample B (n_B = 161)
(Sample A Correlations Below Diagonal; Sample B Above; Decimal Points Omitted)

Test and Subtest	Conventional Subtest								Adaptive Subtest							
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Conventional																
1 VR	--	64	57	27	58	56	64	75	88	64	57	41	49	51	59	71
2 NA	76	--	53	28	43	50	55	62	63	79	51	49	32	53	51	59
3 AR	66	66	--	09	57	65	35	52	52	51	74	30	43	50	37	48
4 CSA	34	32	30	--	29	16	30	37	27	20	22	37	24	15	22	24
5 MR	55	47	53	27	--	71	23	42	52	44	52	33	78	56	22	43
6 SR	58	57	62	25	56	--	21	44	49	47	54	32	56	75	17	41
7 SP	69	70	52	37	36	37	--	69	61	55	37	40	24	25	83	63
8 LU	74	72	55	23	41	44	75	--	75	63	51	48	39	47	66	83
Adaptive																
9 VR	87	70	59	27	48	52	69	73	--	66	55	40	46	50	59	66
10 NA	73	86	66	24	47	58	68	69	74	--	57	47	40	54	52	61
11 AR	59	63	77	24	38	57	49	50	56	67	--	34	44	60	44	50
12 CSA	48	50	38	40	31	27	49	46	49	46	34	--	21	36	35	39
13 MR	52	40	42	27	73	43	28	34	47	43	32	29	--	51	27	44
14 SR	60	65	60	14	45	78	39	48	58	65	64	34	38	--	25	47
15 SP	69	67	47	31	36	41	84	73	67	66	49	47	32	45	--	67
16 LU	68	70	51	25	39	44	69	81	68	68	53	40	34	48	72	--

spect to predictive validity. Model CDPZE, in particular, provided the best summary of the data from this viewpoint. This model had the smallest cross-validation index for Sample B and the second smallest for Sample A. It appears preferable to the CDP model because it is a less restrictive model, not requiring a multiplicative structure for $\mathbb{D}_\zeta$ and $\mathbb{D}^2_\eta$.

Parameter estimates for model CDPZE for both samples are listed in Table 4. Estimated correlations among the abilities are moderate to high in value. The only exceptions are the correlations of the SP test with the MR and SR tests in Sample B. The estimated method correlation between the conventional and adaptive tests across the subtests was very high for both samples (.98 and .97 for samples A and B, respectively). Two notable findings appear consistently in both samples: The scaling fac-

Table 2
Mean (M) and Standard Deviation (SD) for Conventional and Adaptive Subtests of the DAT for Samples A and B

	Sample A				Sample B			
	DAT		CAT		DAT		CAT	
Test	M	SD	M	SD	M	SD	M	SD
VR	23.96	11.53	23.98	10.97	23.22	10.86	23.09	10.88
NA	23.20	8.85	22.43	8.83	22.22	8.85	21.84	8.93
AR	30.02	8.66	30.10	9.14	29.83	8.91	28.65	9.51
CSA	44.02	11.62	48.19	14.65	43.65	13.24	47.64	14.29
MR	44.18	10.76	44.30	10.84	43.01	11.63	43.27	11.08
SR	31.14	11.73	31.08	12.38	30.29	12.62	28.99	12.60
SP	59.80	15.85	62.35	15.47	60.08	16.51	63.57	15.33
LU	26.58	9.55	25.50	10.18	26.08	9.34	25.21	10.37

Table 3
Discrepancy Indices During Calibration (F_{AA}, F_{BB}) and Cross-Validation (F_{AB}, F_{BA}) for Models Fit to Covariance Matrices of DAT and CAT Subscales

Model	Parameters	F_{AA}	F_{AB}	F_{BB}	F_{BA}
Equal Common Score Variances	52	.762	1.930	1.135	1.389
Congeneric	60	.677	1.831	.914	1.423
CDP	47	.676	1.730	.989	1.270*
CDPZ	54	.629	1.744	.915	1.322
CDPZE	61	.561*	1.642*	.784*	1.273

*Denotes smallest value in column.

tor (in $\hat{\mathbf{D}}_\zeta$) for the observed scores on the CSA test is much greater for the adaptive than for the conventional test, and the ratio of the unique-score standard deviation to the common-score standard deviation (in $\hat{\mathbf{D}}_\eta^2$) is greater than unity for the conventional CSA test. These two results imply that there is greater variability among scores on the adaptive CSA test than on the conventional CSA test.

Discussion

The high correlation between the conventional and adaptive versions of the DAT obtained in both samples for the best cross-validating direct-product model, CDPZE, suggests that the two versions of the test battery are very much alike. Under the multiplicative model and its associated restrictions and assumptions, there is strong evidence for the structural similarity of the conventional and adaptive versions of the DAT.

Results obtained for the DAT are similar to those previously reported for the ASVAB (Cudeck, 1985; Moreno et al., 1984). Taken together, the findings suggest that some degree of structural equivalence can be expected when conventional measures of differential abilities are presented in a carefully developed adaptive mode.

Table 4
Parameter Estimates for Model CDPZE for Both Samples

	Diag($\hat{\mathbf{D}}_\zeta$)		Diag($\hat{\mathbf{D}}_\eta^2$)		$\hat{\rho}_t$							
Sample and Test	DAT	CAT	DAT	CAT	1	2	3	4	5	6	7	8
Sample A												
1 VR	11.07	10.14	.30	.41	1.00							
2 NA	8.35	8.23	.36	.39	.84	1.00						
3 AR	7.97	8.03	.45	.56	.74	.80	1.00					
4 CSA	6.45	11.02	1.50	.88	.67	.66	.56	1.00				
5 MR	9.86	8.98	.44	.69	.63	.56	.57	.49	1.00			
6 SR	10.38	11.41	.54	.44	.69	.74	.78	.45	.60	1.00		
7 SP	14.80	13.98	.37	.46	.79	.79	.61	.69	.42	.49	1.00	
8 LU	8.92	8.99	.38	.52	.83	.82	.68	.59	.48	.57	.86	1.00
Sample B												
1 V	10.38	9.88	.27	.42	1.00							
2 NA	8.00	8.03	.48	.47	.75	1.00						
3 AR	8.16	8.38	.49	.56	.68	.69	1.00					
4 CSA	6.93	10.40	1.63	.93	.56	.67	.45	1.00				
5 MR	11.76	9.07	.00[a]	.72	.60	.51	.65	.48	1.00			
6 SR	12.11	10.90	.36	.62	.63	.63	.76	.46	.75	1.00		
7 SP	15.29	13.79	.38	.44	.69	.64	.46	.55	.25	.26	1.00	
8 LU	8.79	8.98	.32	.53	.82	.73	.62	.66	.46	.53	.77	1.00

[a]Boundary condition.

Even though the overall test battery demonstrated a degree of structural equivalence on cross-validation, the findings suggest that the computer-analogue versions of speeded tests (such as the CSA subtest) present problems. This result agrees with Greaud and Green's (1986) conclusions about computerizing such tests. The CSA test is distinct from other tests when administered conventionally (i.e., the uniqueness is very high compared to the common variance); it is much less so when administered by computer.

It is not clear whether the computerized testing mode actually results in measurement of an ability different from that reflected in scores from conventional paper-and-pencil CSA tests. If this were the case, differential prediction of relevant criterion measures would be expected for the conventional and adaptive forms of the tests. In some domains (e.g., vocational placement), scores on the computerized versions of the test may have more relevance for certain criteria (e.g., contemporary workplace demands) than the older, conventional tests.

It is conceivable that the success with which batteries of differential ability tests seem to have been transferred to a computerized adaptive format could result in an enthusiastic effort to develop CAT versions of personality, interest, or attitude tests. However, the unique measurement problems of these domains may be sufficiently different from tests in the ability domain to make optimistic a priori expectations premature. Instead, a cautious approach—CAT versions of single tests, followed later by CAT versions of entire batteries—will clarify the degree to which the CAT tests can be substituted for the conventional paper-and-pencil measures.

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52,* 317–332.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Bennett, G. K., Seashore, H. G., & Wesman, A. G. (1982). *Differential Aptitude Tests Administrator's Handbook*. San Antonio TX: The Psychological Corporation.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588–606.

Braun, H. I., & Holland, P. W. (1982). Observed score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9–49). New York: Academic Press.

Browne, M. W. (1984). The decomposition of multitrait-multimethod matrices by generalized least squares. *British Journal of Mathematical and Statistical Psychology, 37,* 1–21.

Browne, M. W., & Cudeck, R. A. (in press). Single-sample cross-validation indices for covariance structures. *Multivariate Behavioral Research.*

Cudeck, R. (1985). A structural comparison of conventional and adaptive versions of the ASVAB. *Multivariate Behavioral Research, 20,* 305–322.

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin, 105,* 317–327.

Cudeck, R., & Browne, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research, 18,* 147–167.

Greaud, V. A., & Green, B. F. (1986). Equivalence of conventional and computer presentation of speed tests. *Applied Psychological Measurement, 10,* 23–34.

Gulliksen, H. (1968). Methods for determining equivalence of measures. *Psychological Bulletin, 70,* 534–544.

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). *Causal analysis: Assumptions, models and data*. Beverly Hills CA: Sage.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36,* 109–133.

Jöreskog, K. G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika, 43,* 443–475.

Lawley, D. N., & Maxwell, A. E. (1971). *Factor analysis as a statistical method* (2nd ed.). London: Butterworth.

Lord, F. M. (1974). Individualized testing and item characteristic curve theory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Vol. II). San Francisco: Freeman.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale NJ: Erlbaum.

McBride, J. R. (1986, August). *A computerized adaptive edition of the Differential Aptitude Tests*. Paper presented at the meeting of the American Psychological Association, Washington DC.

McBride, J. R., Corpe, V. A., & Wing, H. (1987, August). *Equating the Computerized Adaptive Edition of the Differential Aptitude Tests*. Paper presented at

the meeting of the American Psychological Association, New York.

Moreno, K. E., Wetzel, C. D., McBride, J. R., & Weiss, D. J. (1984). Relationship between corresponding Armed Services Vocational Aptitude Battery (ASVAB) and computerized adaptive testing (CAT) subtests. *Applied Psychological Measurement, 8,* 155–163.

Owen, R. A. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70,* 351–356.

Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology, 19,* 49–57.

Sympson, J. B., Weiss, D. J., & Ree, M. J. (1982). *Predictive validity of conventional and adaptive tests in an Air Force training environment* (Report AFHRL-TR-81-40). Brooks Air Force Base TX: Manpower and Personnel Division.

Tanaka, J. S. (1987). How big is big enough?: Sample size and goodness of fit in structural equation models with latent variables. *Child Development, 58,* 134–146.

Tanaka, J. S., & Huba, G. J. (1985). A fit index for covariance structure models under arbitrary GLS estimation. *British Journal of Mathematical and Statistical Psychology, 38,* 621–635.

U.S. Department of Defense. (1982). *Armed Services Vocational Aptitude Battery*. North Chicago IL: U.S. Military Entrance Processing Command.

Acknowledgments

This research was supported in part by Advanced Education Project and Grant RSP 1031 from IBM, by a National Research Service Award Predoctoral Nurse Fellowship to the first author, and by an Eva O. Miller Fellowship from the Graduate School of the University of Minnesota to the second author.

Author's Address

Send requests for reprints or further information to Susan J. Henly, Department of Psychology, University of Minnesota, 75 East River Road, Minneapolis MN 55455, U.S.A.