A Multivariate Model Sampling Procedure and a Method of Multidimensional Tailored Testing

Vern W. Urry
U.S. Civil Service Commission

It has been established that in the unidimensional case in which only one ability is tested (Jensema, 1974, 1976; Urry, 1974, 1975, 1977), computer-assisted tailored testing can be very economical. The reliabilities typical of paper-and-pencil tests of single abilities can be achieved with far fewer items. This economy results from the computer-interactive adaptation of the item sequence to the level of ability of the particular examinee. In one study (Urry, 1977) the number of items required to achieve a specified reliability was only one-fifth as great as the number required by a conventional paper-and-pencil test: The paper-and-pencil test required 100 questions and the tailored test required only 20.

When several calibrated item banks are available for the measurement of several abilities, the problem becomes one of determining the most economical use of these banks to optimize the multidimensional validity of a composite of the several tailored test scores. The conventional paper-and-pencil test analogue of this problem is to differentially allocate the number of items to the various tests in a battery so that the multiple correlation is maximized for a fixed testing time. This particular problem was addressed earlier by Taylor (1939, 1950), Horst (1949, 1956), and Woodbury and Novick (1968).

The problem is addressed in the present paper from the perspective of tailored testing. The Woodbury and Novick solution is modified to provide--at an asymptotic value of the maximum multiple correlation as a function of testing time--allocations by item banks for (1) terminal reliabilities, (2) terminal standard errors, and (3) appropriate weights for ability estimates obtained from tailored testing. The modified solution can be used in conjunction with the Owen unidimensional algorithm. The result is a multidimensional algorithm appropriate for use when (1) tests are tailored from several item banks for each examinee and (2) an external measure of job proficiency is available.

In the Owen algorithm, tailored testing can be terminated when a specific value of the standard error of the ability estimate is achieved. The specific value for a given bank is referred to as a terminal standard error. In actuality, this is the square root of the Bayesian posterior variance. Since the standard deviation of ability has been set equal to unity, 1.0 minus the Bayesian posterior variance will yield the terminal reliability when tailored testing has been terminated at a specific terminal standard error. A composite score, or predicted criterion score, can then be obtained by merely adding tailored test scores (ability estimates) that have been multiplied by their appropriate weights. The appropriate weights in this context are regressed score weights. The extent of regression, or the standard deviation of the regressed scores, is given by the square root of the terminal reliability. This is the construct validity coefficient or the slope of the regression of tailored test scores (ability estimates) on true ability. True ability, through calibration, has a standard deviation of 1.0.

In the balance of this paper, the modified Woodbury and Novick solution is detailed, a multivariate item response generator is described which generated data for a simulation study, the design of the simulation study to assess the effectiveness of the multidimensional algorithm is presented, results obtained from the simulation study are given, and the important implications of the multidimensional algorithm for tailored testing are considered.

Method

Multivariate Item Response Generation

In multivariate item response generation, several true ability scores and a true criterion score are sampled for each simulated examinee. Given these true ability scores, binary item responses (that is, zeroes or ones indicating incorrect or correct answers) are sampled for the items. For convenience, the item responses for each simulated examinee are arranged on the basis of the particular ability each item measures. These data are then available for the simulation of multidimensional tailored testing. Each ability can be estimated through a unidimensional tailoring algorithm using the appropriate item responses; each ability estimate can then be compared with its corresponding true value. In addition, appropriate weights can be applied to each ability estimate to obtain a composite or predicted criterion score that can, in turn, be compared with a true criterion parameter.

An estimate of the population supermatrix P is required. This symmetric supermatrix has the following partitioned structure:

$$P = \left[\begin{array}{c|c} P_{\theta\theta} & \underline{\rho} \\ \hline \underline{\rho}' & 1.00 \end{array}\right] = \left[\begin{array}{cccc|c} 1.00 & \rho_{\theta_1\theta_2} & \cdots & \rho_{\theta_1\theta_p} & \rho_{\theta_1 y} \\ \rho_{\theta_2\theta_1} & 1.00 & \cdots & \rho_{\theta_2\theta_p} & \rho_{\theta_2 y} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_{\theta_p\theta_1} & \rho_{\theta_p\theta_2} & \cdots & 1.00 & \rho_{\theta_p y} \\ \hline \rho_{y\theta_1} & \rho_{y\theta_2} & \cdots & \rho_{y\theta_p} & 1.00 \end{array}\right] \qquad [1]$$

where $P_{\theta\theta}$ is the matrix of intercorrelations between the latent abilities, θ_k, for $k = 1, 2 \ldots p$;

$\underline{\rho}$ is a column vector of correlations $\rho_{\theta_k y}$ between the latent abilities θ_k and a criterion variable, y; and

$\underline{\rho}'$ is the transpose of $\underline{\rho}$, or a row vector of the validity coefficients for the latent abilities.

While the supermatrix P is never observed in practice,[1] it can be satisfactorily estimated from supermatrices of attenuated correlations based on large samples through the use of

$$P = D_r^{-\frac{1}{2}} \left[R-I\right] D_r^{-\frac{1}{2}} + I. \qquad [2]$$

In Equation 2, the matrix I is the $(p+1)$ by $(p+1)$ identity matrix. The supermatrix R is partitioned as follows:

$$R = \left[\begin{array}{c|c} R_{\hat{\theta}\hat{\theta}} & \underline{r} \\ \hline \underline{r}' & 1.00 \end{array}\right] = \left[\begin{array}{cccc|c} 1.00 & r_{\hat{\theta}_1\hat{\theta}_2} & \cdots & r_{\hat{\theta}_1\hat{\theta}_p} & r_{\hat{\theta}_1\hat{y}} \\ r_{\hat{\theta}_2\hat{\theta}_1} & 1.00 & \cdots & r_{\hat{\theta}_2\hat{\theta}_p} & r_{\hat{\theta}_2\hat{y}} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{\hat{\theta}_p\hat{\theta}_1} & r_{\hat{\theta}_p\hat{\theta}_2} & \cdots & 1.00 & r_{\hat{\theta}_p\hat{y}} \\ \hline r_{\hat{y}\hat{\theta}_1} & r_{\hat{y}\hat{\theta}_2} & \cdots & r_{\hat{y}\hat{\theta}_p} & 1.00 \end{array}\right] \qquad [3]$$

where $R_{\hat{\theta}\hat{\theta}}$ is the matrix of attenuated intercorrelations between less than perfectly reliable or fallible measures of the latent abilities, $\hat{\theta}_k$, for $k = 1, 2 \ldots p$;

$\underline{r}$ is a column vector of attenuated correlations, $r_{\hat{\theta}_k\hat{y}}$, between the fallible measures of latent abilities, $\hat{\theta}_k$, and a fallible criterion variable, $\hat{y}$; and

$\underline{r}'$ is the transpose of $\underline{r}$, or a row vector of validity coefficients attenuated in the variables.

[1] The supermatrix P represents the intercorrelations between perfectly reliable ability and criterion measures. While this supermatrix exists in theory, in practice, perfectly reliable ability and criterion measures are exceptional.

The supermatrix $D_r^{-\frac{1}{2}}$ is partitioned as follows:

$$D_r^{-\frac{1}{2}} = \left[\begin{array}{c|c} D^{-\frac{1}{2}}_{r_{\hat{\theta}\hat{\theta}}} & \underline{0} \\ \hline \underline{0}' & \frac{1}{\sqrt{r_{\hat{y}\hat{y}}}} \end{array}\right] = \left[\begin{array}{ccccc|c} \frac{1}{\sqrt{r_{\hat{\theta}_1\hat{\theta}_1}}} & 0 & \cdots & 0 & & 0 \\ 0 & \frac{1}{\sqrt{r_{\hat{\theta}_2\hat{\theta}_2}}} & \cdots & 0 & & 0 \\ \cdots & \cdots & \cdots & 0 & & \cdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{r_{\hat{\theta}_p\hat{\theta}_p}}} & & 0 \\ \hline 0 & 0 & \cdots & 0 & & \frac{1}{\sqrt{r_{\hat{y}\hat{y}}}} \end{array}\right] \qquad [4]$$

where $D^{-\frac{1}{2}}_{r_{\hat{\theta}\hat{\theta}}}$ is a p by p diagonal matrix of reciprocal square roots of the reliabilities of the fallible measures of the latent abilities;

$\underline{0}$ is a column or p by 1 null vector;

$\underline{0}'$ is the transpose of $\underline{0}$ or a row or 1 by p null vector; and

$\frac{1}{\sqrt{r_{\hat{y}\hat{y}}}}$, a scalar, is the reciprocal square root of the reliability of the criterion variable.

The estimate of the supermatrix P is decomposable into its eigenvectors and eigenvalues. This process yields the identity

$$\hat{\text{P}} = QDQ' \ , \qquad [5]$$

where Q is the $(p+1)$ by $(p+1)$ matrix of the eigenvectors of $\hat{\text{P}}$;

D is the $(p+1)$ by $(p+1)$ diagonal matrix of the eigenvalues of $\hat{\text{P}}$ in descending order of magnitude; and

$Q'$is the transpose of Q.

A $(p+1)$ by $(p+1)$ matrix of weights

$$W = D^{\frac{1}{2}}Q' \qquad [6]$$

is obtained for later use, where the diagonal matrix $D^{\frac{1}{2}}$ contains the square roots of the eigenvalues in descending order of magnitude and the matrix Q' is as previously defined. A matrix T can then be obtained through

$$T = ZW, \qquad [7]$$

where Z is the N by $(p+1)$ matrix,the elements of which are merely independent, drawing from the normal distribution, $N(0,1)$, with a mean of zero and a variance of one; and the matrix W is as defined in Equation 6. The matrix T is partitioned as follows:

$$T = \left[\begin{array}{c|c} \theta & \underline{y} \end{array}\right] = \left[\begin{array}{ccccc|c} \theta_{11} & \theta_{12} & \cdots & & \theta_{1p} & y_1 \\ \theta_{21} & \theta_{22} & \cdots & & \theta_{2p} & y_2 \\ \cdots & \cdots & \cdots & & \cdots & \cdots \\ \cdots & \cdots & \theta_{ij} & & \cdots & y_j \\ \cdots & \cdots & \cdots & & \cdots & \cdots \\ \theta_{N1} & \theta_{N2} & \cdots & & \theta_{Np} & y_N \end{array}\right] \qquad [8]$$

where θ is the N by p matrix of ability parameters in standard score form, and $\underline{y}$ is the N by 1 column vector of criterion parameters in standard score form.

Since the expectation, E, of $\frac{Z'Z}{N}$ is equal to the identity matrix, i.e.,

$$E\left[\frac{Z'Z}{N}\right] = I \ , \qquad [9]$$

it then follows that the expectation, E, of $\frac{T'T}{N}$, or

$$E\left[\frac{T'T}{N}\right] = QD^{\frac{1}{2}}\, E\left[\frac{Z'Z}{N}\right] D^{\frac{1}{2}}Q' = \hat{\mathrm{P}}, \qquad [10]$$

is equal to the desired supermatrix. Thus, the operation in Equation 7 provides a simulated sample of size N from the population defined by $\hat{\mathrm{P}}$.

Let it be assumed that representative values for the item parameters are available for items measuring the several abilities. For example, the i^{th} item measuring the k^{th} ability would have the parameters $\underline{a}_{ik}$ (discriminatory power), b_{ik} (difficulty), and c_{ik} (coefficient of guessing), where $i = 1, 2, \ldots q_k$, the number of items measuring the k^{th} ability; q_k is the last item in the k^{th} bank; and $k = 1, 2, \ldots p$, the number of abilities measured.

Given θ_{jk} (examinee j's true ability score on ability k) and the parameters for item i on ability k, the j^{th} simulated examinee's binary response, or u_{ijk}, (that is, 0 or 1 indicating an incorrect or correct answer) is obtained through evaluating

$$P'_{ik}(\theta_{jk}) = c_{ik} + (1-c_{ik})P_{ik}(\theta_{jk}) \ , \qquad [11]$$

where $P'_{ik}(\theta_{jk})$ is the proportion obtaining a binary score of 1 at θ_{jk}, and

$P_{ik}(\theta_{jk})$, the proportion knowing the correct answer (as opposed to knowing or guessing correctly) at θ_{jk} is

$$P_{ik}(\theta_{jk}) = \left[1 + \exp\{-Da_{ik}(\theta_{jk} - b_{ik})\}\right]^{-1} , \qquad [12]$$

where D is the constant 1.7. Because of the complementary relationship,

$$Q_{ik}(\theta_{jk}) = 1 - P'_{ik}(\theta_{jk}) , \qquad [13]$$

the exhaustive and mutually exclusive events (that is, 0 or 1 indicating an incorrect or correct answer to the i^{th} item measuring the k^{th} ability) may be mapped onto the unit interval.

Thereafter, a random number, r_u, is drawn from a distribution of uniform density on the interval from 0 to 1. Given that

$$r_u > P'_{ik}(\theta_{jk}) , \qquad [14]$$

assign

$$u_{ijk} = 0 \text{ (incorrect)} . \qquad [15]$$

Otherwise, or when

$$r_u \leq P'_{ik}(\theta_{jk}) , \qquad [16]$$

assign

$$u_{ijk} = 1 \text{ (correct)} . \qquad [17]$$

The process is merely repeated for distinct u_{ijk}.

For the purpose of subsequent processing, it is convenient to structure a complete record for the j^{th} simulated examinee in the following manner:

$$\begin{array}{ccccccc}
\theta_{j1} & u_{1j1} & u_{2j1} & \cdots & u_{ij1} & \cdots & u_{q_1 j1} \\
\theta_{j2} & u_{1j2} & u_{2j2} & \cdots & u_{ij2} & \cdots & u_{q_2 j2} \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\theta_{jk} & u_{1jk} & u_{2jk} & \cdots & u_{ijk} & \cdots & u_{q_k jk} \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\theta_{jp} & u_{1jp} & u_{2jp} & \cdots & u_{ijp} & \cdots & u_{q_p jp} \\
y_j & & & & & &
\end{array}$$

where θ_{jk} is the parameter in standard form on the k^{th} ability for the j^{th} examinee;

u_{ijk} is the binary score for the j^{th} examinee on the i^{th} item measuring the k^{th} ability; and

y_j is the criterion parameter in standard form for the j^{th} examinee.

For convenience, the number of the last item in the k^{th} item bank, q_k, is illustrated in this record as constant across banks. It need not be held constant in practice.

Modified Woodbury and Novick Solution

The developments in this portion of the paper parallel those given by Woodbury and Novick (1968), with some notational change. Where simplifications were possible and modifications necessary, these are introducted.

The investigators designate a diagonal matrix D with main diagonal elements d_{kk} given by

$$d_{kk} = \frac{\sqrt{t_k}}{\sqrt{\sigma^2_{x_k(t_k)}\left[1-\rho_{x_k(t_k)x'_k(t_k)}\right]}} \, , \qquad [18]$$

where t_k is the time allocated to the k^{th} test in the battery. The variance and reliability of the k^{th} test are $\sigma^2_{x_k(t_k)}$ and $\rho_{x_k(t_k)x'_k(t_k)}$, respectively, for the allocated time, t_k. Thus, the total time allocated to the n tests in the battery is given by

$$T = \sum_{k=1}^{n} t_k \, . \qquad [19]$$

In further developments, a product of matrices is useful:

$$D\Sigma D = \left[D^* D^{-\frac{1}{2}}_{diag\ \Sigma}\right] \Sigma \left[D^{-\frac{1}{2}}_{diag\ \Sigma}\ D^*\right] = D^*RD^* \, , \qquad [20]$$

where the elements of D were defined in Equation 18, and Σ is the variance-covariance matrix for the tests under the allocated times that sum to T. In actual practice, it will be convenient to work with the righthand equality, where R (the intercorrelation matrix for the tests under the allocated times that sum to T)is given by

$$R = D^{-\frac{1}{2}}_{diag\ \Sigma}\ \Sigma\ D^{-\frac{1}{2}}_{diag\ \Sigma} \, , \qquad [21]$$

as indicated by the equalities in Equation 20 in which $D^{-\frac{1}{2}}_{diag\ \Sigma}$ is a diagonal matrix of reciprocal square roots of the diagonal elements (test variances) of Σ. Since D is the product of D^* and $D_{diag\ \Sigma}^{-\frac{1}{2}}$, D^* may be observed to have diagonal elements

$$d^*_{kk} = \frac{\sqrt{t_k}}{\sqrt{1 - \rho_{x_k(t_k)x'_k(t_k)}}}, \qquad [22]$$

where the terms are as previously defined. The matrix

$$F = D_t^{-1}(D^*RD^* - D_t)\ D_t^{-1} \qquad [23]$$

is then obtained for future use. In this equation, D_t is a diagonal matrix with elements t_k (as previously defined) on its main diagonal. The diagonal matrix D_t^{-1} contains, by convention, the reciprocals of t_k on its main diagonal.

It is also required that there be a column vector

$$D\ \underline{Cov}_{x(t)y} \Big/ \sqrt{\sigma_y^2} = D^* D_{diag\ \Sigma}^{-\frac{1}{2}}\ \underline{r}_{x(t)y}\sigma_y \Big/ \sqrt{\sigma_y^2} = D^*\underline{r}_{x(t)y}, \qquad [24]$$

where D, D^*, $D_{diag\ \Sigma}^{-\frac{1}{2}}$ are as previously defined;

$\underline{Cov}_{x(t)y}$ is a column vector of covariances, $r_{x_k(t_k)y}\ \sigma_{x_k(t_k)}\ \sigma_y$, between the tests, under the allocated times that sum to T, and the criterion y;

σ_y and σ_y^2 are the standard deviation and variance, respectively, for the criterion y; and

$\underline{r}_{x(t)y}$ is a column vector of validity coefficients under the allocated times that sum to T.

Again, in actual practice, it is convenient to work with the righthand equality in Equation 24. For future use, a column vector

$$\underline{\gamma} = D_t^{-1}\ D^*\ \underline{r}_{x(t)y} \qquad [25]$$

is obtained where all terms have been previously defined.

A valid solution requires that all the optimally allocated testing times, t^*_k ,in the column vector

$$\underline{t}^* = \frac{T^* + \underline{e}'\ F^{-1}\ \underline{e}}{\underline{e}'\ F^{-1}\underline{\gamma}.}\ F^{-1}\underline{\gamma} - F^{-1}\underline{e} \qquad [26]$$

be non-negative. In Equation 26, T^* is the total time available for testing, i.e.,

$$T^* = \sum_{k=1}^{n} t_k^* \qquad [27]$$

where the column vector $\underline{e}$ is an elementary vector of length $\underline{n}$, the elements of which are all unity;

$\underline{e}'$, the transpose of $\underline{e}$, is a row vector of length n, the elements of which are all unity;

F^{-1} is the inverse of the output matrix of Equation 23; and

$\underline{\gamma}$ is the output vector of Equation 25.

An application is begun with a large value for T^*, which is then decreased systematically. When a negative t_k^* is encountered, the k^{th} test is dropped from the solution by appropriately reducing the involved matrices, vectors, and scalars. In this case, the terminal reliability, the terminal standard error, and the weight for regressed ability estimate on the k^{th} test are set to zero, unity, and zero, respectively; and the original subscripting is preserved for the purpose of tailored testing.

A diagonal matrix D_{t*} is defined to contain the elements of $\underline{t}^*$ of Equation 26 on its main diagonal. The diagonal matrix then required is

$$\Lambda = \left[I + (D_{t*}\, D_t^{-1} - I)\, D_r\right] D_{t*}^{-1}\, D_t \qquad [28]$$

where I is the identity matrix;

D_{t*}, D_t, D_t^{-1} are as previously defined;

D_r is a diagonal matrix containing the reliabilities of the tests under the allocated times that sum to T on its main diagonal; and

D_{t*}^{-1} is the inverse of D_{t*}.

At a particular T^*, the terminal reliabilities for tailored testing are then given on the main diagonal of

$$D_{\tilde{r}} = \Lambda^{-1} D_r\ , \qquad [29]$$

where Λ^{-1} is the inverse of the diagonal matrix defined in Equation 28. It is of interest to interpret a diagonal element of Equation 29. From Equations 28 and 29, it can be deduced that a diagonal element of $D_{\tilde{r}}$, namely $\tilde{r}_{kk}$, is defined as

$$\tilde{r}_{kk} = \frac{\dfrac{t_k^*}{t_k}\, r_{kk}}{1 + \left(\dfrac{t_k^*}{t_k} - 1\right) r_{kk}}\ , \qquad [30]$$

which is the continuous form of the Spearman-Brown formula, where $\dfrac{t_k^*}{t_k}$ is the continuous analogue of the discrete integer k as given in

$$\tilde{r}_{kk} = \frac{k\ r_{kk}}{1 + (k - 1)\ r_{kk}} \qquad [31]$$

under the usual notation for the formula. The diagonal elements of $D_{\tilde{r}}$ are thus the appropriately altered reliabilities from a solution for a particular T^*.

At a particular T^*, the terminal standard errors for tailored testing are given on the main diagonal of

$$D_{\tilde{\sigma}_{\varepsilon}} = \left[I - D_{\tilde{r}}\right]^{\frac{1}{2}} \quad , \qquad [32]$$

where I is the identity matrix and $D_{\tilde{r}}$ is defined in Equation 29. Thus, a diagonal element of $D_{\tilde{\sigma}_{\varepsilon}}$, namely $\tilde{\sigma}_{\varepsilon_k}$, can be interpreted from Equation 32 as

$$\tilde{\sigma}_{\varepsilon_k} = \sqrt{1 - \tilde{r}_{kk}} \quad , \qquad [33]$$

the square root of 1 minus the reliability.

The squared maximum multiple correlation for the weighted composites of ability estimates, R^2_c, for a particular T^* is given by

$$R^2_c = \underline{r}'_{x(t)y}\ (R + \Lambda - I)^{-1}\ \underline{r}_{x(t)y} \quad , \qquad [34]$$

where $\underline{r}'_{x(t)y}$, a row vector, is the transpose of $\underline{r}_{x(t)y}$ as previously defined in Equation 24;
R and Λ are as previously defined in Equations 20 and 28, respectively; and
I is the identity matrix.

It is necessary to derive the appropriate weights for regressed ability estimates at a given T^*. Standard weights, $\tilde{\beta}^*_k$, are obtained through the normal equations provided by calculus,

$$\underline{\tilde{r}}_{x(t^*)y} = \tilde{R}\underline{\tilde{\beta}}^* \quad , \qquad [35]$$

where $\underline{\tilde{r}}_{x(t^*)y}$ and $\tilde{R}$ are the appropriately altered column validity vector and intercorrelation matrix for the tests in which the allocations of time sum to T^*. The altered column validity vector is known from Woodbury and Novick (1968) to be given by

$$\tilde{\underline{r}}_{\tilde{\underline{x}}(t^*)y} = \Lambda^{-\frac{1}{2}}\,\underline{r}_{\underline{x}(t)y} \quad , \qquad [36]$$

and the altered intercorrelation matrix is similarly known to be given by

$$\tilde{R} = \Lambda^{-\frac{1}{2}}\left[R + \Lambda - I\right]\Lambda^{-\frac{1}{2}} \quad . \qquad [37]$$

Thus, an explicit solution for $\tilde{\underline{\beta}}^*$ involves the premultiplication of both sides of Equation 35 by $\tilde{R}^{-1}$. Then

$$\tilde{\underline{\beta}}^* = \tilde{R}^{-1}\tilde{\underline{r}}_{\tilde{\underline{x}}(t^*)y} \quad . \qquad [38]$$

But it is known that

$$\tilde{R}^{-1} = \Lambda^{\frac{1}{2}}\left[R + \Lambda - I\right]^{-1}\Lambda^{\frac{1}{2}} \quad , \qquad [39]$$

because the inverse of a product of square basic matrices is equal to the product of their inverses in reverse order. Substitution from Equations 36 and 39 into Equation 38 now provides a more convenient form,

$$\tilde{\underline{\beta}}^* = \Lambda^{\frac{1}{2}}\left[R + \Lambda - I\right]^{-1}\underline{r}_{\underline{x}(t)y} \quad . \qquad [40]$$

The weights for regressed ability estimates appropriate in tailored testing are now obtained from

$$\tilde{\underline{b}} = D_{\tilde{r}}^{-\frac{1}{2}}\,\underline{\beta}^* \quad , \qquad [41]$$

because the main diagonal elements of $D_{\tilde{r}}^{-\frac{1}{2}}$ are the reciprocal square roots of the reliabilities or the reciprocal standard deviations of the regressed ability estimates provided by the Owen algorithm. A predicted criterion score, $\hat{y}_j$, is obtained with

$$\hat{y}_j = \hat{\underline{\theta}}'\tilde{\underline{b}} = \hat{\underline{\theta}}'\,D_{\tilde{r}}^{-\frac{1}{2}}\,\tilde{\underline{\beta}}^* \quad , \qquad [42]$$

where $\hat{\underline{\theta}}'$ is a row vector of regressed ability estimates from the Owen algorithm (one for each ability bank) and the remaining terms are as previously defined. The middle equality in Equation 42 is the most convenient form, but the righthand equality is more informative. In the righthand equality, the product $\hat{\underline{\theta}}'\,D_{\tilde{r}}^{-\frac{1}{2}}$ can be seen to standardize ("unregress") the regressed ability estimates prior to the application of standard weights; concomitantly, this product can be viewed as an operation that unbiases the regressed ability estimates or renders them on the same scale as the corresponding true abilities.

Of interest, in actual practice, are the asymptotic properties of the maximum multiple correlation as T^* increases. Beyond some point on T^*, increased testing time yields diminishingly small increases in validity, as indexed by the maximum multiple correlation. A solution at a specific T^*

is then selected in which negligible increase in the maximum multiple correlation is expected with an increase in testing time. The terminal reliabilities, terminal standard errors, and appropriate weights for tailored testing may then be obtained for the selected value of T^*.

Design of the Simulation Study

The population intercorrelation matrix, by assumption, contains the intercorrelations of the latent abilities, their correlations with a criterion, and the criterion self-correlation, unity. To assure verisimilitude, this matrix of parameters should be an intercorrelation matrix actually obtained in a large sample and later disattenuated in the tests. A matrix of parameters implies error-free tests. An attenuated matrix was obtained from French (1963). This matrix is well known because it was also used in the six-predictor-variable problem analyzed in the original Woodbury and Novick article. The reliabilities required to disattenuate this matrix appropriately were .76, .82, .70, .64, and .74.

The particular population matrix used to generate the intercorrelated true ability and criterion parameters, given in Table 1, is partitioned. The last row and column of the matrix contains the true validity vector and the criterion self-correlation, unity. The larger partitioned area contains the intercorrelations between the latent or true abilities.

Table 1
The Assumed Population Matrix

	True Abilities						Criterion
Variable	1	2	3	4	5	6	7
1	1.00	.16	.47	.38	.46	.33	.50
2	.16	1.00	.22	.12	.21	.09	.17
3	.47	.22	1.00	.30	.53	.34	.51
4	.38	.12	.30	1.00	.49	.21	.36
5	.46	.21	.53	.49	1.00	.36	.50
6	.33	.09	.34	.21	.36	1.00	.20
7	.50	.17	.51	.36	.50	.20	1.00

Through the use of this matrix, 900 simulated cases were randomly sampled, each case having six true ability parameters, θ_{jk},for k = 1, 2, . . . 6 and one criterion parameter, y_j. The subscript j indexed the simulated cases where j = 1, 2, . . . N.

In order to generate the item responses for the simulated cases for each item (ability) bank or test, the item parameters must be specified. The distinction between item bank and test depends on the use of the particular simulated case as described below. For convenience, each bank or test had the same parameter specifications. The parameters for the 100 items in each bank or test were specified as follows: In sequence, 20 items were assigned to each level of a_i, viz., .8, 1.2, 1.6, 2.0, and 2.4 that in turn contained 20 levels of b_i varying from -1.9 to 1.9 in increments of .2. The c_i were successively assigned the values from .03 to .27 in increments of .03, where both c_1 and c_{100} were accordingly .03. Again, to assure verisimilitude, these specifications were in accord with reasonable expectations for ability test items.

For 500 of these simulated cases, the generated item responses were scored as they would be for six 100-item conventional tests. Raw scores--merely the number of items answered correctly--were obtained for the 100-item test variables. The scores were then intercorrelated along with the criterion, Kuder-Richardson Formula 20 reliabilities were estimated, and a modified Woodbury and Novick solution was obtained.

Using the obtained solution, multidimensional tailored testing was then conducted with the 400 simulated cases remaining from the original 900. These cases were evenly divided into two samples, viz., Sample 1 and Sample 2. For each case, tailored testing proceeded by using each bank until the particular terminal standard error, $\tilde{\sigma}_{\varepsilon_k}$, was achieved. The tailored test scores, or ability estimates, were then weighted to obtain $\hat{y}_j$, the predicted or estimated criterion parameter. These estimates were then correlated with their corresponding and known true parameters in order to allow an assessment of the effectiveness of the multidimensional algorithm.

Results

The 100-item tests for the first 500 simulated cases were conventionally scored. Their reliabilities were computed by means of Kuder-Richardson Formula 20, and the tests were intercorrelated along with the criterion. The results are reported in Table 2. The off-diagonal elements of this matrix should resemble the off-diagonal elements of the assumed population matrix (as given in Table 1) from which the 500 simulated cases were sampled in order to allow the generation of item responses. The resemblance is unmistakable because both the sampling error for the 500 cases and the measurement error, as indicated by the high test reliabilities, were small.

Table 2

Obtained Reliabilities (Main Diagonal), Test Intercorrelations, and Validity Coefficients (Last Row and Column, Omitting the Main Diagonal), N=500

Variable	1	2	3	4	5	6	7
1	(.963)	.171	.510	.378	.422	.353	.512
2	.171	(.962)	.181	.122	.157	.083	.171
3	.510	.181	(.960)	.317	.516	.373	.469
4	.378	.122	.317	(.961)	.436	.221	.287
5	.422	.157	.516	.436	(.958)	.333	.487
6	.353	.083	.373	.221	.333	(.958)	.206
7	.512	.171	.469	.287	.487	.206	1.000

Using the data given in Table 2, the modified Woodbury and Novick procedure was applied. For this application, the initial testing times for the tests, t_k, were all set to unity. Accordingly, the diagonal matrix D_t was the identity matrix.

The pertinent results obtained in this application are summarized in Table 3. In the notation of this table, the squared construct validity coefficient, $\tilde{\rho}^2_{\hat{\theta}_k \theta_k}$, has been substituted for its identity, $\tilde{r}_{kk}$, the terminal reliability. This identity is readily established by merely correcting the reliability coefficient for attenuation in one of the parallel forms to obtain the validity coefficient. This coefficient is the correlation between the attenuated fallible parallel form and the disattenuated parallel form representing the errorless latent ability or pertinent psychological construct. Squaring this construct validity coefficient merely removes the radical, again yielding the reliability coefficient. This identity assumes that true score is a linear function of true ability.

In Table 3 terminal reliabilities ($\tilde{\rho}^2_{\hat{\theta}_k \theta_k}$), terminal standard errors ($\tilde{\sigma}_{\varepsilon_k}$), and the regressed estimate weights ($\tilde{b}_k$) are given for each bank at five levels of testing time (T^*) along with their associated maximum multiple correlations (R). It should be noted that the maximum multiple correlation increased with testing time and that these increases diminished in magnitude as testing time increased; the maximum multiple correlation as a function of testing time (T^*) eventually reached an asymptotic value beyond which further testing time (T^*) yielded no further return in validity (R). It should also be noted that relatively large increases in the terminal reliabilities were required for negligible increases in validity (R) as testing time (T^*) increased.

In this context, the terminal standard errors are completely determined by the terminal reliabilities. These are merely the square root of 1.0 minus the particular reliability. It is of interest to note that the banks with the higher terminal reliabilities, $\tilde{\rho}^2_{\hat{\theta}_k \theta_k}$, also have the larger regressed estimate weights, $\tilde{b}_k$. The ordering is perfect.

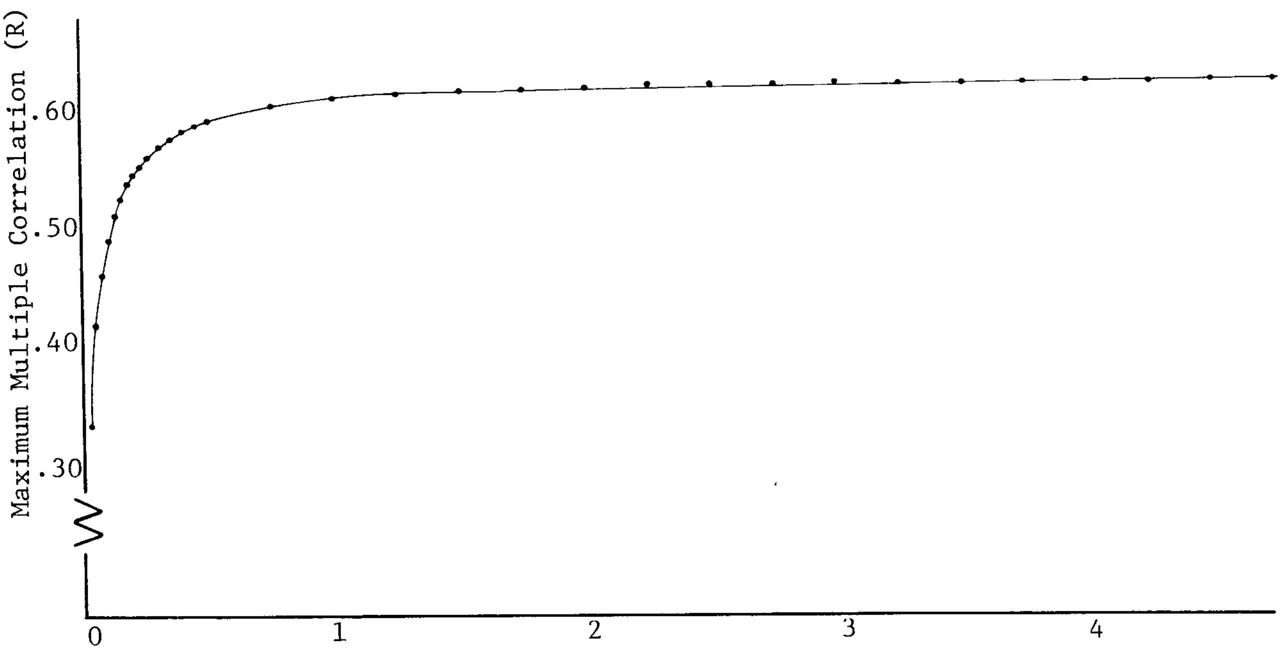
The asymptotic properties of the maximum multiple correlation (R) as a function of testing time (T^*) may be readily observed in Figure 1. Here this function is given for testing times (T^*) of zero through five. There is an abrupt rise in this function in the range of T^* of zero through one; thereafter, increases in the maximum multiple correlation tended to be negligible. The asymptote of the function is approximately .61. As a result, it was decided to use the modified solution at a T^* of 1, where the maximum multiple correlation was approximately .60. This solution yielded the terminal standard errors ($\tilde{\sigma}_{\varepsilon_k}$) and regressed estimate weights ($\tilde{b}_k$) for the six item

Table 3

Terminal Reliabilities ($\tilde{\rho}^2_{\hat{\theta}_k\theta_k}$), Terminal Standard Errors ($\tilde{\sigma}_{\varepsilon_k}$), and Regressed Estimate Weights ($\tilde{b}_k$) for Six Item (Ability) Banks at Five Levels of Testing Time (T*), and the Associated Maximum Multiple Correlation (R)

	Item Bank (k)					
Testing Time (T*)	1	2	3	4	5	6
1 (R=.598)						
$\tilde{\rho}^2_{\hat{\theta}_k\theta_k}$	.913	.440	.835	.000	.892	.000
$\tilde{\sigma}_{\varepsilon_k}$	.295	.748	.406	1.000	.329	1.000
$\tilde{b}_k$	.328	.052	.180	.000	.282	.000
2 (R=.607)						
$\tilde{\rho}^2_{\hat{\theta}_k\theta_k}$	.953	.690	.909	.000	.942	.000
$\tilde{\sigma}_{\varepsilon_k}$	.216	.557	.301	1.000	.241	1.000
$\tilde{b}_k$	.320	.049	.171	.000	.275	.000
3 (R=.610)						
$\tilde{\rho}^2_{\hat{\theta}_k\theta_k}$	.968	.786	.938	.000	.960	.000
$\tilde{\sigma}_{\varepsilon_k}$	.178	.463	.250	1.000	.199	1.000
$\tilde{b}_k$	.317	.048	.168	.000	.272	.000
4 (R=.612)						
$\tilde{\rho}^2_{\hat{\theta}_k\theta_k}$	.976	.835	.952	.000	.970	.000
$\tilde{\sigma}_{\varepsilon_k}$	.155	.406	.218	1.000	.173	1.000
$\tilde{b}_k$	.315	.047	.167	.000	.270	.000
5 (R=.613)						
$\tilde{\rho}^2_{\hat{\theta}_k\theta_k}$	.980	.867	.962	.000	.976	.000
$\tilde{\sigma}_{\varepsilon_k}$	.140	.365	.196	1.000	.156	1.000
$\tilde{b}_k$	.314	.046	.166	.000	.270	.000

Figure 1
The maximum multiple correlation (R) as a function of testing time (T*)



Testing Time (T*)

(ability) banks (k) that are indicated in columns 1, 2, and 3 in Table 4. If the square roots of the terminal reliabilities given in Table 3 for T^* equal to 1 are taken, they will represent forecasts that can be made relative to the $\rho_{\hat{\theta}_k \theta_k}$, or the constant validity of tailored test scores, where the criterion is the particular latent ability. These forecasts, the $\rho_{\hat{\theta}_k \theta_k}$, are given in row 5 of Table 4. The solution selected after the application of the modified procedure forecasts a cross-validity, $\tilde{\rho}_{yy}$, of .60 allowing, of course, for no shrinkage; this forecast is shown in row 4. It should be noted that the abilities measured by Banks 4 and 6 were not required in criterion performance. The terminal standard errors $\tilde{\sigma}_{\varepsilon_4}$ and $\tilde{\sigma}_{\varepsilon_6}$, both 1.00 for Banks 4 and 6, imply that $\tilde{\rho}_{\theta_4 \theta_4}$ and $\tilde{\rho}_{\theta_6 \theta_6}$ both equal zero. Thus, unidimensional tailored testing was unnecessary with respect to Banks 4 and 6.

Multi-bank tailored testing was then conducted using the 400 simulated cases remaining from the original 900. These cases were evenly divided into two samples, viz., Sample 1 and Sample 2. For each case, tailored testing proceeded by using each bank until the particular terminal standard error, $\tilde{\sigma}_{\varepsilon_k}$, was achieved. The four tailored test scores, the $\hat{\theta}_{jk}$, were then weighted to obtain $\hat{y}_j$, the predicted criterion score. The obtained correlations can be directly compared with the forecasts of theory provided in Table 4. These comparisons indicate that the multidimensional procedure performed very well

Table 4
Multidimensional Tailored Testing:
Forecasted and Obtained Results

		Item Bank (k)						
Statistic		1	2	3	4	5	6	
Terminal Standard Error	$\tilde{\sigma}_{\varepsilon_k}$	.30	.75	.41	1.00	.33	1.00	
Regressed Score Weight	$\tilde{b}_k$	.33	.05	.18	.00	.28	.00	
Forecasted Results:								
Validity	$\tilde{\rho}_{\hat{\theta}_k \theta_k}$	.95	.66	.91	.00	.94	.00	
Cross-Validity	$\tilde{\rho}_{yy}$							.60
Simulation Results:								
Sample 1								
Validity	$r_{\hat{\theta}_k \theta_k}$	.95	.70	.91	.00	.95	.00	
Average No. Items	$\bar{m}$	7.20	1.00	4.20	.00	6.10	.00	
Total No. Items	$\Sigma\bar{m}$							18.50
Cross-Validity	r_{yy}							.59
Sample 2								
Validity	$r_{\hat{\theta}_k \theta_k}$	.94	.68	.92	.00	.94	.00	
Average No. Items	$\bar{m}$	7.10	1.00	4.00	.00	6.00	.00	
Total No. Items	$\Sigma\bar{m}$							18.10
Cross-Validity	r_{yy}							.64

in terms of the reduction in the total of the average number of items required per examinee vis-a-vis the number of items typically used in conventional paper-and-pencil test batteries.

Discussion

When external criterion measures are available, the economy of multidimensional tailored testing derives from

1. low values for the terminal reliabilities,
2. a reduced number of measured abilities, and
3. an allocation of terminal reliabilities to minimize computer-interactive time.

Low values for the terminal reliabilities are possible because the maximum multiple correlation (validity) as a function of testing time typically approaches asymptotioally high values quite rapidly. High values for the terminal reliabilities, requiring larger numbers of items, are necessary only after this function has approached the asymptote.

Abilities that are not valid or those sufficiently well measured through correlated abilities are not measured. Thus, correlated abilities pose no particular problem. A proper economy in items is observed even when abilities are correlated. If the items measuring different abilities vary systematically in average examinee response time, computer-interactive time is minimized through the use of the appropriate initial time matrix D_t in obtaining a modified solution. These initial testing times need only be proportional to attain a proper solution.

Whether a modified solution is found for conventional tests or tailored tests, it is equally applicable later in multidimensional tailored testing. The case in which the modified solution was found through conventional testing and later applied in multidimensional tailored testing was illustrated in this paper. The modified solution can also be found for tailored tests and then applied in multidimensional tailored testing.

Certain salient aspects of tailored testing will now be considered by means of the Owen (1969, 1975) algorithm. It is well known that the correlation coefficient completely determines the correlational surface. Thus, a function completely determined by the correlation coefficient will do likewise. For example, if the standard error of the estimate, $\sqrt{1 - \rho^2_{\hat{\theta}_k\theta_k}}$ (the error about the regression of θ_k on $\hat{\theta}_k$), is controlled by the appropriate termination of tailored testing, the slope of this regression, $\rho_{\hat{\theta}_k\theta_k}$, is also controlled. Since the correlational surface implies equality and symmetry of errors about both regression lines, the error about the regression of $\hat{\theta}_k$ on θ_k, which is more traditionally considered to be the standard error of measurement, is equal to the standard error of the estimate, $\sqrt{1 - \rho^2_{\hat{\theta}_k\theta_k}}$ (In this context, the traditional distinction between these standard errors, drawn from classical test theory, breaks down algebraically.)

This determination of the correlational surface also implies a marginal distribution of $\hat{\theta}_k$ that has a scaling identical to that of the marginal distribution of θ_k, true ability, where the mean is zero and the standard deviation is unity. To allow this feature of control over the correlational surfaces,

1. tailored testing with the Owen algorithm must begin with zero and unity for the prior estimates of ability and the standard error of the estimate, respectively; and
2. the scaling of a mean of zero and a standard deviation of unity for θ_k must have been employed when the item parameters were being estimated in large random samples from the population of interest.

Appropriate termination requires variable-length tailored tests to control the resulting correlational surfaces. Fixed-length tailored tests do not provide this control because the standard error of the estimate is, out of necessity, ignored. Evidence of the effectiveness of this control through variable-length tailored testing is provided in Table 4, as well as in several other studies (Urry, 1974, 1975, 1977).

In this context, fixed-length tailored tests clearly result in curvilinear regressions of estimated ability on the corresponding true ability because the standard error of the estimate is ignored. Equiprecision of measurement throughout the important range of each ability is lost when, as in fixed-length tailored testing, the standard error of the estimate is ignored. Terminating the variable-length tailored sequences after a specific value of the standard error of the estimate has been achieved guarantees equiprecision of measurement for the important range of each ability.[2]

Through the multidimensional algorithm, a conditional maximum multiple correlation or validity coefficient at a fixed testing time is sought. To attain the unconditional maximum validity coefficient, infinite testing time would be required. While this is clearly the solution, it has a severe practical drawback in requiring an infinite number of items. Fortunately, increases in testing time beyond some point yield negligible returns with respect to the validity coefficient. There is, then, a trade-off involved in which the appropriate testing time can be rigorously established in a decision theoretic framework. For example, a specific cost/benefit ratio can uniquely determine the appropriate testing time. In this context, the computer could be used to monitor (1) validities, (2) costs, and (3) benefits. Thus would be known on a continual basis (1) the validities of the procedures for personnel selection; (2) the dollar costs of obtaining pertinent test information; and (3) the dollar benefits in increased productivity accruing from the selection decisions.

Preliminary work in decision or utility theory has already been initiated at the U. S. Civil Service Commission. The findings indicate that the dollar benefits of personnel testing tend to be grossly underestimated by both practitioner and sponsor. A close examination of the value of personnel testing would afford a realistic reappraisal, which is much needed after the controversy surrounding personnel testing during the passing decade.

References

French, J. W. The validity of new tests for the performance of college students with high-level aptitude (Research Bulletin 63-7). Princeton, NJ: Educational Testing Service, 1963.

[2] In a Bayesian context, the standard error of the estimate is the proper term to use in determining equiprecision. The reciprocal square root of the information function is appropriate only in a maximum likelihood context. Error reduction is more rapid in the Bayesian context and occurs to a greater extent when incorrect answers are encountered. This can be deduced from Equation 3.7d provided by Owen (1975, p. 353). Greater efficiency in the Bayesian context can result in a correlation between the length of variable-length tailored tests and ability estimates because examinees of lower ability provide more incorrect answers. Hence, fewer itmes are required.

Horst, P. Determination of optimal test length to maximize the multiple correlation. _Psychometrika_, 1949, _14_, 79-88.

Horst, P. Optimal test length for maximum differential prediction. _Psychometrika_, 1956, _21_, 51-66.

Jensema, C. J. The validity of Bayesian tailored testing. _Educational and Psychological Measurement_, 1974, _34_, 757-766.

Jensema, C. J. Bayesian tailored testing and the influence of item bank characteristics. _Proceedings of the first conference on computerized adaptive testing_ (PS-75-6, U. S. Civil Service Commission,Personnel Research and Development Center). Washington DC: U.S. Government Printing Office, 1976. (Superintendent of Documents Stock No. 006-000-00940-9)

Owen, R. J. _A Bayesian approach to tailored testing_ (Research Bulletin 69-92). Princeton, NJ: Educational Testing Service, 1969.

Owen, R. A. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. _Journal of the American Statistical Association_, 1975, _70_, 351-356.

Taylor, C. W. _A method of combining tests into a battery in such a fashion as to maximize the correlation with a given criterion for any fixed total time of testing_. Unpublished master's thesis, University of Utah, 1939.

Urry, V. W. _Computer-assisted testing: Calibration and evalution of the verbal ability bank_ (Technical Study 74-3). Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center, December 1974.

Urry, V. W. _Computer-assisted testing with live examinees: A rendezvous with reality_ (Technical Research Note 75-3). Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center, January 1975.

Urry, V. W. A five-year quest: Is computerized adaptive testing feasible? _Proceedings of the first conference on computerized adaptive testing_ (PS-75-6, U.S. Civil Service Commission, Personnel Research and Development Center). Washington, DC: U.S. Government Printing Office, 1976. (Superintendent of Documents Stock No. 006-000-00940-9)

Urry, V. W. Tailored testing: A successful application of latent trait theory. _Journal of Educational Measurement_, 1977, _14_, 181-196.

Woodbury, M. A., & Novick, M. R. Maximizing the validity of a test battery as a function of relative test lengths for a fixed total testing time. _Journal of Mathematical Psychology_, 1968, _5_, 242-259.