Measuring Test Compromise in High-Stakes Computerized Adaptive Testing: A Bayesian Strategy for Surrogate Test-taker Detection

Daniel O. Segall
Defense Manpower Data Center
Monterey Bay, CA

Abstract

This paper presents a new method for assessing consistency of test performance across two occasions, where on one occasion the level of performance may be misrepresented, and on the second occasion it is not. The new procedure is based on the application of Bayesian model assessment methodology to multidimensional item response theory. A simulation study based on a high-stakes multiple-aptitude test-battery was conducted to evaluate the proposed method. The new procedure was judged superior to one based only on a discriminant analysis of final test-scores.

1. Introduction

For many employers and institutions of higher education, a natural tension exists between the desire to attract the most able applicants and the unwelcome requirement of a stressful high-stakes screening exam. In many cases, the exam used to enforce minimum qualification standards is believed to discourage highly qualified applicants, and in this sense may serve to reduce the number of able candidates. Consequently, test-developers have sought ways of making the testing process less onerous and burdensome from the test-takers' perspective. The move towards computer-administered and computerized-adaptive testing by many test-developers is a direct outcome of these efforts.

Computer administered tests possess a number of advantages over conventional paper-and-pencil tests. The use of a computer allows individually administered standardized tests which can be tailored to the characteristics or needs of the test-taker. The difficulty-level of test questions can be tailored to the test-taker's aptitude level, resulting in increased measurement efficiency (in the case of Computerized Adaptive Testing). Similarly, scheduling

This paper was presented at the annual meeting of the National Council on Measurement in Education (April, 2001), Seattle, WA. The views expressed are those of the author and not necessarily those of the Department of Defense, or the United States government.

Requests for copies should be sent to: Daniel O. Segall, Defense Manpower Data Center, DoD Center Monterey Bay, 400 Gigling Road, Seaside, CA 93955-6771. Email: publications@danielsegall.com

(of the exam's time and place) can be more easily adapted to individual preferences than is possible with conventional group administered paper-and-pencil exams.

Some consideration has been given to in-home (or internet) based computer administered tests as a means to further increase test-takers' comfort and convenience, and ultimately as a way to increase their participation rates in mandatory testing programs. One impediment to in-home high-stakes testing involves the integrity of the reported test performance. In the absence of an impartial proctor, how can the institution verify that the answers attributed to one individual were not in fact provided by another (presumably more able) test-taker? Similarly, how can the institution verify that inappropriate resources (e.g., dictionaries and encyclopedias) were not used as aids by the test-taker in the unmonitored privacy of their home? In a typical high-stakes setting, these assurances are provided by proctors who verify the examinee's identity and adherence to test-security protocol. With unproctored in-home exams, some test-takers may be tempted to enlist the aid of others to achieve high scores and all the benefits derived from these improved scores.

One solution to the verification dilemma requires the administration of a short second exam given under secure proctored conditions. If performance levels on the initial (unproctored/in-home) exam are consistent with the short proctored verification exam, then the in-home scores become the applicant's scores of record, otherwise the initial test-scores are invalidated and the applicant is required to retake an alternate form of the full-length exam under proctored conditions.

The usefulness of the verification approach depends in large part on the efficiency and accuracy of the verification test. Ideally, the short accurate verification test would identify those misrepresenting their performance, while placing only a small additional burden on the honest test-taker. Such a short accurate verification test has the potential to increase participation rates among prospective applicants. Conversely, there are likely to be smaller increases in participation rates resulting from longer less efficient verification exams, since these exams place additional burdens and discomfort on the test-taker.

This paper presents a new method for assessing consistency of test performance across two occasions, where on one occasion the level of performance may be enhanced or misrepresented, and on the second occasion it is not. The new procedure is based on the application of Bayesian model assessment techniques to item response theory. The performance of the proposed method is evaluated through the use of simulated data based on a high-stakes multiple aptitude test-battery.

2. Bayesian Model Comparison

The problem of detecting deceitful test-takers can be approached through the application of Bayesian model comparison methodology (O'Hagan, 1994, Chapter 7). We begin by letting T and V denote mutually exclusive item-sets administered under *initial* and *verification* conditions, respectively, and let $A = T \cup V$ denote the total set of administered items. Initial items are assumed to be administered under conditions that allow for the possibility of misrepresented test-performance (i.e., as is possible with in-home exams). In contrast, verification items are assumed to be administered under conditions that do not allow misrepresentation to occur (because of the presence of proctors and the enforcement of standardized testing practices). The source of administered test items is denoted by $h = (h_i : i \in A)$, where $h_i = 1$ indicates that $i \in T$ (i.e., item i is contained in the initial test), and $h_i = 0$ indicates that $i \in V$ (i.e., item i is contained in the verification test). Responses to the n test items are denoted by the n-element vector $u = (u_i : i \in A)$, where $u_i = 1$ if item i is answered correctly, and $u_i = 0$ otherwise.

We also assume that each test-taker can be classified into one of two mutually exclusive classes: *authentic* (a test-taker who does not misrepresent their performance) or *enhanced* (a test-taker who does misrepresent their performance). The behavior of enhanced test-takers is dependent on item-source (initial or verification), whereas the behavior of authentic test-takers is the same for both item-sources. The behavior of authentic test-takers on both initial and verification items (spanning m dimensions) is described by Model 0 (denoted by $\alpha = 0$), which characterizes the examinee's performance in terms of a vector of true ability parameters $\theta^{(0)} = (\theta_1, ..., \theta_m)$. The test-taking behavior of enhanced examinees is described by Model 1 (denoted by $\alpha = 1$), which is dependent on item source. The performance of enhanced test-takers on initial items is characterized by an enhanced ability vector $\theta^{(1)} = (\theta_1, ..., \theta_m, \theta_{m+1}, ..., \theta_{2m})$, where performance levels along each of the m-dimensions are indexed by summed parameters $\theta_k + \exp(\theta_{k+m})$ (for $k = 1, ..., m$). Here, $\exp(\theta_{k+m})$ is viewed as a positive performance increment. For each dimension k ($k = 1, ..., m$) these increments can result from the substitution of the test-taker (with ability θ_k) with a more able surrogate [possessing enhanced ability $\theta_k + \exp(\theta_{k+m})$]. In contrast, the performance of enhanced test-takers on verification test items ($i \in V$) is modeled as a function of true ability parameters θ_k (for $k = 1, ..., m$) only, since verification items are assumed to be administered under secure testing conditions.

2.1 Posterior Model Probability

From Bayes' theorem, the posterior probability[1] that $\alpha = 1$ (i.e., the initial test performance was misrepresented) given item response data u is:

$$p(\alpha = 1|u) = p(\alpha = 1, u) / p(u) \ , \tag{1}$$

where

$$p(u) = p(\alpha = 0, u) + p(\alpha = 1, u) \ . \tag{2}$$

The joint probability terms of u and α are given by

$$p(u, \alpha = 0) = p(\alpha = 0) \, p(u|\alpha = 0) \tag{3}$$

where

$$p(u|\alpha = 0) = \int \cdots \int p(u|\theta^{(0)}, \alpha = 0) p(\theta^{(0)}|\alpha = 0) \ d\theta_1 \cdots d\theta_m \ , \tag{4}$$

and by

$$p(u, \alpha = 1) = p(\alpha = 1) \, p(u|\alpha = 1) \tag{5}$$

[1] For notational simplicity, we denote different distributions in the same equation or expression by p, with $p(\cdot|\cdot)$ denoting a conditional probability density and $p(\cdot)$ denoting a marginal density. In each instance, the arguments are dependent on the context.

where

$$p\,(u|\alpha = 1) = \int \cdots \int p(u|\theta^{(1)}, \alpha = 1)p(\theta^{(1)}|\alpha = 1)\,d\theta_1 \cdots d\theta_{2m}\ . \tag{6}$$

Here $p\,(\alpha = 1) = 1 - p\,(\alpha = 0)$ are known prior probabilities, and $p(\theta^{(0)}|\alpha = 0)$ denotes a multivariate normal density with an m-element mean vector μ_0 and an $m \times m$ covariance matrix Φ_0. Similarly, we model $p(\theta^{(1)}|\alpha = 1)$ by a $2m$-variate normal density with mean vector μ_1 and covariance matrix Φ_1, where the first m elements of μ_1 are constrained to equal those of μ_0, and the upper left $m \times m$ quadrant of Φ_1 is constrained to equal Φ_0. These constraints follow from the fact that the first m elements of $\theta^{(1)}$ are equal to the m-element vector $\theta^{(0)}$.

The probability of response pattern u for fixed $\theta^{(\alpha)}$ and Model α (for $\alpha = 0, 1$) is calculated from the product of terms associated with individual items:

$$p(u|\theta^{(\alpha)}, \alpha) = \prod_{i \in A} p(u_i = 1|\theta^{(\alpha)}, \alpha)^{u_i}[1 - p(u_i = 1|\theta^{(\alpha)}, \alpha)]^{1-u_i}\ ,$$

which follows from the standard item response theory assumption of local independence. For $\alpha = 0$, the conditional probability of a correct response to the ith item is given by an expanded version of the three-parameter logistic model with item specific discrimination, difficulty, and guessing parameters a_i, b_i, and c_i, respectively:

$$\begin{aligned} P_i(\theta^{(0)}|\alpha = 0) &\equiv p(u_i = 1|\theta^{(0)}, \alpha = 0) \\ &= c_i + \frac{1 - c_i}{1 + \exp[-Da_i'(\theta^{(0)} - b_i 1_m)]}\ , \end{aligned} \tag{7}$$

where $D = 1.7$, b_i is a scalar difficulty-parameter, c_i is the guessing parameter, 1_m is a $m \times 1$ vector of 1's, and a_i' is a $1 \times m$ vector of item discrimination parameters. For $\alpha = 1$, the conditional probability of a correct response to the ith item is given by

$$\begin{aligned} P_i(\theta^{(1)}|\alpha = 1) &\equiv p(u_i = 1|\theta^{(1)}, \alpha = 1) \\ &= c_i + \frac{1 - c_i}{1 + \exp(-Da_i' J_i \tilde{\theta}_i)}\ , \end{aligned} \tag{8}$$

where $\widetilde{\theta}_i = \{\theta_1 - b_i, ..., \theta_m - b_i, \exp(\theta_{m+1}), ..., \exp(\theta_{2m})\}$, $J_i = (I_m, h_i I_m)$ is an $m \times 2m$ partitioned matrix, and where I_m is an $m \times m$ identity matrix. Note that the conditional probability of a correct response for item i under the enhanced model $p(u_i = 1|\theta^{(1)}, \alpha = 1)$ is either equal to (when $i \in V$) or higher (when $i \in T$) than the corresponding probability under the authentic model $p(u_i = 1|\theta^{(0)}, \alpha = 0)$. This follows from noting that the exponent containing h_i in (8) in effect increments the true ability parameters θ_k by positive values $\exp(\theta_{k+m})$ for the enhanced model when $h_i = 1$ (i.e., when the item is an initial item), and does not provide an increment when $h_i = 0$ (i.e., when the item is a verification item).

2.2 Adaptive Item-Selection

Regardless of how items are selected for the initial test, the certainty regarding α can be enhanced though the efficient choice of verification-test items. This can be accomplished through the adaptive administration of items during the verification test phase. Rather than selecting items to minimize the uncertainty regarding latent ability parameters (which is the objective of classical adaptive item selection, van der Linden & Pashley, 2000), a potentially more efficient approach selects items that minimize the posterior uncertainty of α given data u.

One common characterization of posterior uncertainty is provided by the posterior variance:

$$\mathrm{Var}\left(\alpha|u\right) = p\left(\alpha = 1|u\right)\left[1 - p\left(\alpha = 1|u\right)\right] \ . \tag{9}$$

Suppose n items have already been administered, and the corresponding responses are contained in the n-element vector u. Further suppose that we consider each candidate item $j \notin A$ for possible administration as the $(n+1)$-th item, and select the item j' which minimizes the expected posterior variance

$$j' = \min_{j \notin A} \mathrm{E}\left[\mathrm{Var}\left(\alpha|u_j, u\right)\right] \ , \tag{10}$$

where the expectation is taken over the yet-to-be-observed response u_j. This expectation is calculated from the preposterior distribution of u_j given the already observed n responses u:

$$\begin{aligned} \mathrm{E}\left[\mathrm{Var}\left(\alpha|u_j, u\right)\right] &= \mathrm{Var}\left(\alpha|u_j = 1, u\right) p\left(u_j = 1|u\right) \\ &\quad + \mathrm{Var}\left(\alpha|u_j = 0, u\right) p\left(u_j = 0|u\right) \end{aligned} \tag{11}$$

where $p\left(u_j|u\right) = p\left(u_j, u\right)/p\left(u\right)$. The required terms on the right-hand side of (11) can be computed from a straightforward generalization of (1) and (2) which replaces u with the augmented response vectors $(u_j = 0, u)$ and $(u_j = 1, u)$.

Additional items can be selected and administered until either: (a) the posterior variance $\mathrm{Var}\left(\alpha|u\right)$ based on administered items becomes sufficiently small, or (b) some prespecified target test-length has been reached.

2.3 Marginal Probability Approximations

For moderate to high dimensionality problems (moderate to large m), the integration in (4) and (6) can be computationally burdensome using standard numerical quadrature techniques, since the number of function evaluations increases exponentially with m. However, useful approximations can be obtained by the exact integration of a truncated Taylor series expansion. For higher-dimensionality problems, this approach requires far fewer computations than required by numerical quadrature techniques.

Approximations to the marginal probabilities given by (4) and (6) can be obtained through a second-order Taylor series expansion (Tanner, 1996, p. 31) about the posterior mode

$$\hat{\theta}^{(\alpha)} = \max_{\theta^{(\alpha)}} \left[\frac{p(u|\theta^{(\alpha)}, \alpha) p(\theta^{(\alpha)}|\alpha)}{\int p(u|\theta^{(\alpha)}, \alpha) p(\theta^{(\alpha)}|\alpha) d\theta^{(\alpha)}} \right] \ .$$

This expansion takes the general form

$$\begin{aligned} p\,(u|\alpha) &= \int \exp[l_\alpha(\theta^{(\alpha)}|u)]d\theta^{(\alpha)} \\ &= \int \exp[l_\alpha(\hat{\theta}^{(\alpha)}|u) + (\theta^{(\alpha)} - \hat{\theta}^{(\alpha)})' S_\alpha(\hat{\theta}^{(\alpha)}|u) \\ &\quad -\frac{1}{2}(\theta^{(\alpha)} - \hat{\theta}^{(\alpha)})' I_\alpha(\hat{\theta}^{(\alpha)}|u)(\theta^{(\alpha)} - \hat{\theta}^{(\alpha)}) + r_\alpha(\theta^{(\alpha)}|u)]d\theta^{(\alpha)} \\ &\approx p(u|\hat{\theta}^{(\alpha)}, \alpha)p(\hat{\theta}^{(\alpha)}|\alpha)\left|I_\alpha(\hat{\theta}^{(\alpha)}|u)\right|^{-1/2}(2\pi)^{(\alpha+1)m/2} \qquad (12) \end{aligned}$$

where $l_\alpha(\theta^{(\alpha)}|u) = \ln[p(u|\theta^{(\alpha)}, \alpha)p(\theta^{(\alpha)}|\alpha)]$,

$$I_\alpha(\hat{\theta}^{(\alpha)}|u) = -\left.\frac{\partial^2}{\partial(\theta^{(\alpha)})^2} l_\alpha(\theta^{(\alpha)}|u)\right|_{\theta^{(\alpha)}=\hat{\theta}^{(\alpha)}},$$

and $S_\alpha(\hat{\theta}^{(\alpha)}|u) = \partial l_\alpha(\theta^{(\alpha)}|u)/\partial\theta^{(\alpha)}\Big|_{\theta^{(\alpha)}=\hat{\theta}^{(\alpha)}}$. Since the first term $l_\alpha(\hat{\theta}^{(\alpha)}|u)$ is a constant, it can be moved outside the integral. The second term vanishes since the vector of first derivatives evaluated at the mode is equal to zero: $S_\alpha(\hat{\theta}^{(\alpha)}|u) = 0$. Then by ignoring the higher-order terms $r_\alpha(\theta^{(\alpha)}|u)$, the remainder resembles the exponent of a multivariate normal density function with integral given by Anderson (1984, pp. 15–17). The approximation provided by (12) can be calculated from the following steps:

1. Compute the posterior mode $\hat{\theta}^{(\alpha)}$.
2. Compute the determinant of $\left|I_\alpha(\hat{\theta}^{(\alpha)}|u)\right|$ (the determinant of the observed information matrix evaluated at the posterior-mode $\hat{\theta}^{(\alpha)}$).
3. Compute the product $p(u|\hat{\theta}^{(\alpha)}, \alpha) \times p(\hat{\theta}^{(\alpha)}|\alpha)$ (the product of the likelihood and prior evaluated at the posterior-mode) and combine with other terms as indicated in (12) to produce the approximation to $p\,(u|\alpha)$.

Additional computational details are provided in the Appendix.

3. Simulation Study

A simulation study was conducted to examine the performance of the proposed Bayesian detection method when applied to a high-stakes multiple aptitude battery. Responses from examinees assumed to follow the authentic and enhanced ability models were generated for two types of exams: an adaptive initial exam and an adaptive verification exam. For the verification exam, two approaches to item selection and classification were studied. The first approach used traditional adaptive item-selection which maximized the precision of the latent ability parameters. This approach, termed score-based approach, used an optimally weighted linear combination of the resulting initial and verification test-scores to classify respondents. The second approach was based on the Bayesian model-comparison strategy, which used an adaptive item-selection strategy to minimize the posterior uncertainty of α, and a classification index based on the posterior probability of α given all initial and verification item response data u.

Table 1: CAT-ASVAB characteristics

		Test	Pool	a	Battery	
Subtest	Content Area	Length	Size	80th %-tile	Full	Partial
1	General Science (GS)	15	110	1.4	√	
2	Arithmetic Reasoning (AR)	15	209	1.4	√	√
3	Word Knowledge (WK)	15	228	1.7	√	√
4	Paragraph Comprehension (PC)	10	88	1.4	√	√
5	Auto Information (AI)	10	104	1.7	√	
6	Shop Information (SI)	10	103	1.4	√	
7	Math Knowledge (MK)	15	103	2.1	√	√
8	Mechanical Comprehension (MC)	15	103	1.2	√	
9	Electronics Information (EI)	15	97	1.2	√	

Table 2: Covariance matrix of latent abilities Φ

	Dimension								
Dimension	GS	AR	WK	PC	AI	SI	MK	MC	EI
GS	1.000								
AR	.645	1.000							
WK	.908	.611	1.000						
PC	.808	.847	.880	1.000					
AI	.486	.332	.326	.349	1.000				
SI	.676	.424	.566	.514	.824	1.000			
MK	.564	.846	.516	.711	.150	.218	1.000		
MC	.739	.758	.644	.800	.623	.725	.625	1.000	
EI	.808	.639	.724	.743	.642	.724	.536	.822	1.000

3.1 Full and Partial Test Batteries

All item responses were simulated from items patterned after unidimensional subtests contained in the Computerized Adaptive Testing version of the Armed Services Vocational Aptitude Battery ([CAT-ASVAB]; Segall & Moreno, 1999). These subtests, and their disattenuated correlations Φ are listed in Tables 1 and 2, respectively. Two conditions corresponding to two different batteries were simulated: (a) a full battery, and (b) a partial battery. The full battery condition consisted of all $m = 9$ subtests/dimensions spanned by the CAT-ASVAB. The partial battery condition consisted of only $m = 4$ dimensions—the math and verbal portions of the battery. (See the last two columns of Table 1.) These four subtests play a special role in Military selection: A weighted sum of scores from these four tests is used to determine eligibility status for entrance into the Military.

3.2 Simulated Test-taker Characteristics

Ability parameters for two types of examinees were generated. The first type consisted of 2,500 authentic respondents whose responses were consistent with the authentic-ability model ($\alpha = 0$). For these examinees m-element vectors of parameters $\theta^{(0)}$ were specified. The second type consisted of 2,500 enhanced respondents whose responses were consistent with the enhanced-ability model ($\alpha = 1$). For these examinees, $2m$-element vectors of parameters $\theta^{(1)}$ were specified.

For the 2,500 simulated test-takers whose responses were consistent with the authentic-ability model ($\alpha = 0$), the m-element vectors of parameters $\theta^{(0)}$ were drawn from a normal distribution with mean 0, and covariance matrix Φ (shown in Table 2). For the Full-Battery condition ($m = 9$), the entire 9×9 Φ-matrix was used in the covariance matrix specification. For the Partial-Battery condition ($m = 4$), only variances/covariances involving the four math/verbal dimensions (Table 1, last column) were used in specifying the required 4×4 Φ-matrix.

Under the enhanced-ability model, we denote the full vector of ability parameters by

$$\theta_1, ..., \theta_m, \theta_{m+1}, ..., \theta_{2m} \ ,$$

where the first m-elements denote the test-taker's true ability on each of the m latent dimensions, and the second set of m-elements ($m + 1, ..., 2m$) denote corresponding log-increments. These parameters were generated using an approach based on a surrogate test-taker strategy:

1. First, draw a vector of true-abilities for the target test-taker from the population distribution : $(\theta_1, ..., \theta_m) \sim \mathrm{N}(0, \Phi)$.

2. Let $\lambda = (\lambda_1, ..., \lambda_m)$ denote a vector of indicator variables, where $\lambda_k = 1$ if a surrogate with higher ability on dimension k has been identified, and $\lambda_k = 0$ otherwise. Initially $\lambda_k = 0$ (for $k = 1, ..., m$).

3. Sample a surrogate (denoted by s) from the same distribution: $(\theta_1^s, ..., \theta_m^s) \sim \mathrm{N}(0, \Phi)$. If need be, repeatedly sample surrogates until one is found that has higher latent ability levels on all four key[2] dimensions (corresponding to subtests AR, WK, PC, and MK), then assign log ability-increment parameters to the target test-taker in the following manner:

$$\left.\begin{array}{rcl} \theta_{k+m} & = & \ln(\theta_k^s - \theta_k) \\ \lambda_k & = & 1 \end{array}\right\} \text{for all } k \text{ where } \theta_k^s > \theta_k \ .$$

4. For the Full-Battery condition, if any dimensions have not been assigned a surrogate increment-parameter (i.e., if $\sum_k \lambda_k \neq m$), sample a new surrogate and perform the assignments given by

$$\theta_{k+m} = \ln(\theta_k^s - \theta_k) \, , \text{ for all } k \text{ where } (\theta_k^s > \theta_k \cap \lambda_k = 0) \ ,$$

and

$$\lambda_k = 1, \text{ for all } k \text{ where } \theta_k^s > \theta_k \ .$$

Note that increment-parameters are only assigned for previously unassigned dimensions. The surrogate sampling is repeated until all dimensions have been assigned a log-increment parameter.

[2] Scores based on these dimensions are among the most important to the test-taker, since a composite based on these subtests directly influences entrance into the Military. These dimensions are also among the most highly correlated, and are believed to be highly g loaded, as suggested by their high predictive validity for success in military training.

Table 3: Summary statistics of raw gain scores.

	Dimension								
Statistic	GS	AR	WK	PC	AI	SI	MK	MC	EI
Mean	1.19	1.20	1.11	1.25	1.02	1.07	1.15	1.20	1.18
SD	.81	.81	.82	.76	.84	.85	.83	.84	.84
Skewness	.82	.91	.95	.86	1.15	1.08	.95	.81	.82
Kurtosis	.43	.81	.78	.84	1.26	1.06	.82	.35	.46

This algorithm was used to generate ability parameters for the 5,000 simulees assumed to follow the enhanced-ability model. The first m-elements of the prior mean vector μ_1 were set to zero, and the upper-left $m \times m$ submatrix of Φ_1 was set equal to Φ (shown in Table 2). The remaining elements of μ_1 and Φ_1 were specified from sample means and covariances of the $\theta^{(1)}$ parameters generated from 10,000 additional replications of the surrogate test-taker sampling algorithm. The average Pearson product moment correlation r between log-gain and ability across the relevant dimensions was

$$\frac{1}{m^2}\sum_{j=1}^{m}\sum_{k=1}^{m} r\left(\ln\left(\theta_j^s - \theta_j\right), \theta_k\right) = \begin{cases} -0.25, & \text{for Full-Battery } (m = 9) \\ -0.30, & \text{for Partial-Battery } (m = 4) \end{cases},$$

suggesting that lower ability simulees tended to receive slightly larger gains. The average correlation among log-gain variables was

$$\frac{2}{m\,(m-1)}\sum_{k=2}^{m}\sum_{j=1}^{k-1} r\left(\ln\left(\theta_j^s - \theta_j\right), \ln\left(\theta_k^s - \theta_k\right)\right) = \begin{cases} +0.25, & \text{for Full-Battery } (m = 9) \\ +0.33, & \text{for Partial-Battery } (m = 4) \end{cases},$$

suggesting that those receiving large gains on one dimension tended to receive slightly larger gains on other dimensions as well. The first four moments of the marginal distributions of raw $(\theta_k^s - \theta_k)$ gain-scores resulting from 10,000 full-battery[3] replications of the surrogate test-taker sampling algorithm are displayed in Table 3.

3.3 Initial Test

An initial test was simulated for each respondent. Item selection and scoring algorithms, as well as item-pool composition were based on the CAT-ASVAB (Segall, Moreno, Bloxom, & Hetter, 1997; Segall, Moreno, & Hetter, 1997). Item responses were generated according to an m-dimensional *multi-unidimensional* model. According to this approach, each item possessed one nonzero discrimination parameter contained in the m-dimensional vector $a_i' = (a_{1i}, a_{2i}, ..., a_{mi})$. The pattern of nonzero item discrimination parameters for the ith item was dependent on the content (or dimension) k (for $k = 1, ..., m$) measured by the item. The discrimination parameter $a_{ki} > 0$, if item i measures dimension k, and $a_{ki} = 0$ otherwise. The b_i, c_i, and nonzero a_{ki} were defined from the estimated unidimensional three-parameter logistic item parameters from Form 1 of the CAT-ASVAB. These

[3] Although the sample moments of the 10,000 *partial-battery* replications are not displayed, they were within sampling error of their full-battery counterparts—a result assured by the data generation algorithm.

estimated parameters were treated as true (error-free) values for response generation, and for item selection and scoring.

In addition to the characteristics of the items (as indicated by their a, b, and c parameters), simulated responses to the adaptively selected items were also dependent on the examinee classification: authentic respondent ($\alpha = 0$) or enhanced respondent ($\alpha = 1$). For authentic respondents, the conditional probability of a correct response was calculated from (7) which is dependent on the test-taker's ability level $\theta^{(0)}$. For enhanced respondents, the conditional probability of a correct response to items contained in the initial test pool ($h_i = 1 : i \in T$) is calculated from (8), which is dependent on the set of surrogate ability parameters (expressed as functions of the parameters contained in $\theta^{(1)}$). In both instances, dichotomous (correct/incorrect) responses were produced by comparing the relevant conditional response probabilities to pseudo-random uniform numbers.

Using this approach, multidimensional responses can be used to simulate outcomes associated with m separately administered and scored unidimensional adaptive tests (Segall, 1996). This approach is useful in situations where the dimensions spanned by a collection of unidimensional tests are correlated. The number of separately tailored unidimensional adaptive tests simulated for each test-taker depended on the condition. For the Full-Battery condition, nine adaptive tests were simulated. For the Partial-Battery condition, four adaptive tests were simulated. (See Table 1.) Items were adaptively selected to maximize Fisher information (Lord, 1980, pp. 72–73) evaluated at the provisional ability estimate (Owen, 1975). Tests were terminated after a fixed number of administered-items (either 10 or 15; see Table 1), and a final score $\hat{\theta}_k$ for each dimension k was specified as the mode of the posterior distribution. The Bayesian scoring algorithms assumed unidimensional standard normal prior distributions.

3.4 Verification Test

Two verification-testing approaches were simulated. The first was a *score-based* approach, where test-takers were classified on the basis of weighted linear combinations of initial and verification test-scores. The second approach was based on the Bayesian model-comparison strategy.

3.4.1 Score-Based Approach.

Full and partial battery verification tests were simulated using the same conventions as the initial tests described above: maximum information item selection, fixed length tests, Owen's (1975) provisional ability estimator, and final posterior mode scoring. Full and partial battery item pools for verification tests were cloned from the initial test pools[4] summarized in Table 1. For response generation, the source of the administered items was assumed to be from the verification pool $i \in V$ so that $h_i = 0$. As indicated by (7) and (8) (when $h_i = 0$), generated responses to these items were dependent only on the test-taker's true ability level, and were not influenced by the ability level of surrogates, regardless of the test-taker classification. Four different verification tests were simulated: (a) full-battery long (27 items), (b) full-battery short (9 items), (c) partial-battery long (55 items), and (d) partial battery short (20 items). Table 4 provides test-lengths for individual subtests.

[4] Note that there were no restrictions on the administration of items across the initial and verification test item pools. That is, they were assumed to be mutually exclusive for the sake of the simulation study.

Table 4: Verification test-lengths for score-based conditions

	Full Length Battery		Partial Length Battery	
Subtest	Long	Short	Long	Short
GS	3	1		
AR	3	1	15	5
WK	3	1	15	5
PC	3	1	10	5
AI	3	1		
SI	3	1	15	5
MK	3	1		
MC	3	1		
EI	3	1		
Total	27	9	55	20

The result of this verification test was a vector of m scores (posterior modes) for each simulated respondent. These scores were combined with the m initial test scores for the 5,000 respondents to produce discriminant function scores (Fisher, 1936) for each respondent:

$$\hat{y} = X\left(X'X\right)^{-1}X'y,$$

where X denotes a $5000 \times (2m+1)$ matrix where the jth row (denoted by x_j) contains the constant 1 and the $2m$ test-scores of the jth respondent:

$$x_j = (1, \underbrace{\hat{\theta}_{j1}, ..., \hat{\theta}_{jm}}_{\text{Initial}}, \underbrace{\hat{\theta}_{j1}, ..., \hat{\theta}_{jm}}_{\text{Verification}}),$$

and where y denotes a 5000-element vector of true respondent classification

$$y_j = \begin{cases} 1, & \text{if } \alpha = 1 \text{ for the } j\text{th respondent} \\ 0, & \text{otherwise.} \end{cases}$$

The accuracy levels of the predicted classification scores $\hat{y}$ for the full and partial-battery conditions were compared to those produced by the Bayesian method.

3.4.2 Bayesian Surrogate Test-Taker Detection Strategy.

Two hypothetical item pools were constructed for the multidimensional Bayesian approach: one for the full-battery verification test, and another for the partial-battery verification test. The item pool for the full-battery verification test consisted of 117 items. Each item was assumed to load on a single dimension, with 13 items loading on each of the nine dimensions. If for example the ith item loaded on the third (WK) dimension, its pattern of discrimination parameters resembled $a_i' = (0, 0, a_{3i}, 0, 0, 0, 0, 0, 0)$. All items loading on the same dimension were assumed to have equivalent non-zero discrimination parameters. These were set to the 80th-%tile value of the empirical distribution of discrimination parameters (for the unidimensional pools of the corresponding dimension). These are listed in the last column of Table 1. Within each set of 13 items, difficulty values b's were equally spaced from -1.5 to $+1.5$. All guessing parameters c's were set to 0.2.

The item pool for the partial-battery verification test was constructed using similar conventions. It consisted of 116 items; each item loaded on a single dimension, with 29 items loading on each of the four dimensions. All items loading on the same dimension were assumed to have equivalent non-zero discrimination parameters which were set equal to the 80th-%tile value of the empirical distribution of discrimination parameters. (See Table 1). Within each set of 29 items, difficulty values b's were equally spaced from -1.5 to $+1.5$ and all guessing parameters c's were set to 0.2.

For each respondent, the posterior probability (1) was computed from the response data provided by the administered items. The prior classification probability was assumed to be known: $p\left(\alpha=1\right) = p\left(\alpha=0\right) = .5$. The verification test was terminated if the maximum test-length had been reached (30 items for the full-battery condition; 60 items for the partial-battery condition), or if the posterior stopping criterion

$$\min\left[1-p\left(\alpha=1|u\right), p\left(\alpha=1|u\right)\right] < 0.001$$

has been satisfied. This criterion can be equivalently expressed in terms of the posterior variance, where testing continued until

$$\mathrm{Var}\left(\alpha=1|u\right) < 0.001 \times 0.999 \; .$$

If neither test-termination criteria had been satisfied, additional items were selected and administered. These items were chosen to minimize the expected posterior variance (10). Responses generated for selected items were dependent only on the test-taker's true ability level, and were not influenced by the ability level of surrogates, regardless of the test-taker classification as indicated by (7) and by (8) when $i \in V \; (h_i = 0)$.

For long initial tests, item selection computations can be abbreviated somewhat by skipping the multidimensional Bayes modal estimation. Rather than computing four different posterior modes for each candidate item (for each model and possible response), good results can be obtained by evaluating the Taylor series approximation (12) at the posterior modes based on the complete set of administered items (excluding the candidate item). Using this simplification, the posterior modes are computed (and thus updated) only after the item is chosen, and a response is generated. This simplification was used in the simulated item-selection algorithm.

3.5 Results

Receiver operating characteristic (ROC) curves of signal detection theory (Green & Swets, 1966) can be used to examine the relative performance of alternate classification procedures. For a given classification procedure, a point along the ROC curve can be calculated from the two proportions:

$x(t)$: *false-alarm rate*, the proportion of authentic examinees with index values greater than cutoff t, (i.e., the proportion of authentic examinees improperly identified as enhanced), and

$y(t)$: *hit rate*, the proportion of enhanced examinees with index values greater than cutoff t, (i.e., the proportion of enhanced examinees correctly identified as enhanced).

Procedures can be compared on the basis of selected $[x(t), y(t)]$ pairs obtained for various cutoff-values of t applied to the index values.

Classification accuracy rates were calculated for alternative combinations of maximum test-length and posterior variance targets using intermediate results from the simulated tests. The outcomes for the full and partial battery conditions are detailed in Tables 5 and 6, respectively. For each condition, test-length summaries and false-alarm rates for different conditions defined by maximum test-length (Max. n) and posterior variance (PV) stopping rules are displayed. Results for the fixed-length score-based procedure are provided at the top of Tables 5 and 6, while results for the variable-length Bayesian procedure are provided in the main body of each table.

For the full-battery condition, the most accurate classification was provided by the condition depicted in the first row (under Bayesian procedure) which consisted of a maximum test-length of 30 items, with a posterior stopping criterion of $.001 \times .999$. Here we see that the average test length $\overline{n}$ for the authentic test-takers ($\alpha = 0$) was 25.6 items, and 19.8 items for the enhanced-ability test-takers ($\alpha = 1$). We also see that the proportion of authentic test-takers with test-lengths equal to zero $p(n = 0)$ was about 4-percent, compared to the near zero-percent for enhanced test-takers. The false-alarm rates were .01, .02, and .07 for hit rates of .90, .95, and .98, respectively. These are superior to the false-alarm rates of the score-based procedure (based on 27 items) depicted in the first row of Table 5.

For fixed maximum test-length and posterior stopping criteria, the partial-battery condition displayed lower classification-accuracy than the full-battery condition. However, high hit and low false-alarm rates were observed for many stopping-rule combinations. For example, for the condition defined by 20-item maximum test-length and $.001 \times .999$ posterior-variance stopping criteria, the false-alarm rates were .02 and .07 for hit rates of .90 and .95, respectively. In this condition average test-lengths were 19.9 and 14.2 items for authentic and enhanced test-takers, respectively.

The results displayed in Tables 5 and 6 suggest several other notable trends.

I. For both the full and partial-battery conditions, the Bayesian procedure:

   (a) Provides high hit-rates with low false-alarm rates for a number of conditions defined by alternative test termination criteria.

   (b) Outperforms the score-based procedure for conditions of comparable test-lengths.

II. The following notable full-battery outcomes (Table 5) were achieved by the Bayesian method:

   (a) It provides a hit-rate of .98 while achieving a very low false-alarm rate of .07. This level of detection is likely to be satisfactory for high-stakes testing programs where accurate-detection is required.

(b) Over one-fourth of the authentic test-takers [$p(n = 0) = .29$] required no verification test even for the condition which produced a hit-rate of .95 with a false-alarm rate of .06.

(c) Nearly all enhanced test-takers [$p(n > 0) = .94$] required at least some verification test items.

(d) Very short tests (15 items or less) can achieve a high hit-rate (.95) with a low false-alarm rate (.08).

III. The following notable partial-battery outcomes (Table 6) were achieved by the Bayesian method:

(a) It provides a hit-rate of .95 while achieving a false-alarm rate of .03.

(b) Over one-fourth of the authentic test-takers [$p(n = 0) = .29$] required no verification test even for the condition which produced a hit-rate of .90 with a false-alarm rate of .05.

(c) Very short tests (10 items or less) can achieve a high hit-rate (.90) with a low false-alarm rate (.05).

4.0 Discussion

The full-battery simulation study demonstrates that the Bayesian procedure can accurately classify 98 percent of enhanced test-takers, while only misclassifying about 7 percent of the authentic test-takers. In practice, this accuracy level is likely to be satisfactory for even high-stakes tests, where cheating is most likely to provide large payoffs. Results also suggest that test-takers are likely to experience little additional testing burden by verification testing based on the Bayesian approach. The performance of the Bayesian procedure is especially impressive in comparison to the more traditional score-based procedure. With comparable test-lengths, the Bayesian approach provides false-alarm rates that are significantly lower than those produced by the score-based procedure.

Although the accuracy of the Bayesian procedure reported here is high, its accuracy is likely to be situation specific. First, classification accuracy is likely to be influenced by the size and precision of the item pools (both initial and verification). Large pools with highly discriminating items and heterogeneous difficulty parameters are likely to provide the most accurate classification. Second, the size of ability increments (i.e., difference between the surrogate and test-taker's ability levels) are also likely to play prominent roles in determining classification accuracy—with larger average increments providing higher accuracy. Third, the dimensionality of the battery and the length of the initial exam appear to have an effect on classification accuracy: Longer initial-exams spanning many dimensions appear to result in greater accuracy than short initial-exams spanning a smaller number of dimensions.

Another consideration which may limit the generality of the reported findings centers on test-taker behavior with regard to seeking and utilizing surrogate examinee knowledge. In practice, test-takers may use reference materials instead of surrogates, or use a combination of their own knowledge, and knowledge obtained from surrogates and reference materials. Although these scenarios were not explicitly modeled in the simulation study,

Table 5: Full-battery false-alarm rates (proportion of authentic test-takers classified as enhanced test-takers) and test-length statistics for score-based and Bayesian procedures.

		Test-length Statistics				False		
Termination Criteria		Model				Alarm Rates		
		$\alpha = 0$		$\alpha = 1$		Hit-Rate		
Max. n	PV	$p(n = 0)$	$\overline{n}$	$p(n = 0)$	$\overline{n}$	.90	.95	.98
		Score-Based Procedure						
27	—	.00	27	.00	27	.02	.05	.13
9	—	.00	9	.00	9	.10	.20	.39
		Bayesian Procedure						
30	$.001 \times .999$	.04	25.6	.00	19.8	.01	.02	.07
	$.003 \times .997$	.16	20.6	.01	17.0	.01	.03	.14
	$.006 \times .994$	.29	16.2	.03	14.7	.01	.06	.77
	$.009 \times .991$	.37	13.3	.06	13.1	.02	.57	.77
25	$.001 \times .999$	.04	21.8	.00	17.9	.01	.03	.08
	$.003 \times .997$	.16	17.7	.01	15.5	.01	.04	.15
	$.006 \times .994$	.29	14.1	.03	13.5	.01	.07	.77
	$.009 \times .991$	.37	11.7	.06	12.1	.03	.58	.77
20	$.001 \times .999$	.04	17.8	.00	15.6	.01	.05	.13
	$.003 \times .997$	.16	14.7	.01	13.7	.01	.06	.24
	$.006 \times .994$	.29	11.8	.03	12.0	.02	.08	.77
	$.009 \times .991$	.37	9.9	.06	10.8	.04	.58	.77
15	$.001 \times .999$	.04	13.7	.00	12.8	.02	.08	.20
	$.003 \times .997$	.16	11.4	.01	11.4	.02	.08	.33
	$.006 \times .994$	.29	9.3	.03	10.2	.03	.15	.77
	$.009 \times .991$	.37	7.8	.06	9.3	.06	.58	.77
10	$.001 \times .999$	.04	9.3	.00	9.3	.06	.18	.35
	$.003 \times .997$	.16	7.9	.01	8.6	.07	.21	.42
	$.006 \times .994$	.29	6.5	.03	7.8	.09	.28	.77
	$.009 \times .991$	.37	5.6	.06	7.3	.13	.60	.77
5	$.001 \times .999$	.04	4.7	.00	4.9	.16	.31	.53
	$.003 \times .997$	.16	4.1	.01	4.8	.17	.34	.57
	$.006 \times .994$	.29	3.4	.03	4.5	.19	.39	.77
	$.009 \times .991$	.37	3.0	.06	4.3	.24	.61	.77

Table 6: Partial-battery false-alarm rates (proportion of authentic test-takers classified as enhanced test-takers) and test-length statistics for score-based and Bayesian procedures.

		Test-length Statistics				False		
Termination Criteria		Model				Alarm Rates		
		$\alpha = 0$		$\alpha = 1$		Hit-Rate		
Max. n	PV	$p(n = 0)$	$\overline{n}$	$p(n = 0)$	$\overline{n}$	.90	.95	.98
		Score-Based Procedure						
55	—	.00	55	.00	55	.01	.06	.19
20	—	.00	20	.00	20	.05	.12	.31
		Bayesian Procedure						
55	.001 × .999	.00	54.2	.00	25.1	.01	.03	.10
	.010 × .990	.01	39.2	.00	16.9	.01	.03	.10
	.050 × .950	.12	17.7	.01	10.6	.02	.04	.37
	.100 × .900	.29	9.3	.04	6.9	.03	.27	.64
40	.001 × .999	.00	39.6	.00	21.1	.01	.03	.12
	.010 × .990	.01	30.7	.00	14.4	.01	.03	.12
	.050 × .950	.12	14.9	.01	9.3	.02	.04	.36
	.100 × .900	.29	8.2	.04	6.4	.03	.27	.64
30	.001 × .999	.00	29.8	.00	18.0	.01	.04	.16
	.010 × .990	.01	24.5	.00	12.5	.01	.04	.16
	.050 × .950	.12	12.9	.01	8.3	.02	.06	.35
	.100 × .900	.29	7.3	.04	5.9	.03	.28	.64
20	.001 × .999	.00	19.9	.00	14.2	.02	.07	.18
	.010 × .990	.01	17.6	.00	10.1	.02	.07	.18
	.050 × .950	.12	10.3	.01	7.1	.03	.08	.34
	.100 × .900	.29	6.3	.04	5.2	.05	.28	.64
10	.001 × .999	.00	10.0	.00	8.8	.05	.13	.34
	.010 × .990	.01	9.5	.00	6.8	.05	.13	.34
	.050 × .950	.12	6.7	.01	5.2	.06	.14	.40
	.100 × .900	.29	4.5	.04	4.1	.08	.36	.66
5	.001 × .999	.00	5.0	.00	4.9	.14	.26	.50
	.010 × .990	.01	4.9	.00	4.3	.14	.26	.50
	.050 × .950	.12	3.9	.01	3.5	.15	.27	.53
	.100 × .900	.29	2.8	.04	3.0	.16	.42	.68

the multidimensional model may still provide satisfactory detection with an appropriate choice of prior distribution on ability and gain parameters. Further research would be needed to verify this possibility.

In any given application, mis-specification of the prior distribution of ability/log-increments (indexed by mean vector μ_1, and covariance matrix Φ_1) is likely to degrade the performance of the Bayesian procedure—with the amount of degradation related to the degree of mis-specification. In the study described here, distribution parameters involving log-increments were specified on rational grounds by assuming particular behavioral patterns among examinees. In practice, these strong assumptions may not hold—at the very least they should be empirically verified. A more accurate prior specification might be achieved by estimating the full distribution of ability/log-increment parameters directly from the population of interest. In principal, this can be done through an extension of a direct estimation procedure of the sort recommended by Mislevy (1984), where μ_1 and Φ_1 are estimated directly from initial and verification test item responses provided by a group of representative test-takers. Given an accurate specification of the prior, studies of the sort presented here can be conducted to estimate the expected classification accuracy of the proposed method with given item pools and test termination criteria.

Appendix

Modal values $\hat{\theta}^{(\alpha)}$ can be obtained through an iterative numerical procedure such as the Newton-Raphson procedure, where the $(j+1)$-th approximation is given by

$$\theta_{j+1}^{(\alpha)} = \theta_j^{(\alpha)} + I_\alpha^{-1}(\theta_j^{(\alpha)}|u) S_\alpha(\theta_j^{(\alpha)}|u) \ , \tag{13}$$

and where $I_\alpha(\theta_j^{(\alpha)}|u)$ and $S_\alpha(\theta_j^{(\alpha)}|u)$ denote the first and minus second derivatives evaluated at $\theta_j^{(\alpha)}$ (see below). In some instances, convergence can be improved by replacing $I_\alpha^{-1}(\theta_j^{(\alpha)}|u)$ by its expected value, which can be obtained by substituting $P_i(\theta^{(\alpha)}|\alpha)$ for u_i in (14) below. Successive iterations are obtained until convergence has been achieved, usually indicated when $\theta_{j+1}^{(\alpha)} \approx \theta_j^{(\alpha)}$.

First Derivatives

The required first derivatives are given by Segall (1996, p. 340, Equation 25):

$$S_0(\theta^{(0)}|u) \equiv \frac{\partial}{\partial \theta^{(0)}} l_0(\theta^{(0)}|u) = \sum_{i \in A} v_i(\theta^{(0)}|\alpha = 0) a_i - \Phi_0^{-1}(\theta^{(0)} - \mu_0) \ ,$$

and

$$S_1(\theta^{(1)}|u) \equiv \frac{\partial}{\partial \theta^{(1)}} l_1(\theta^{(1)}|u) = \sum_{i \in A} v_i(\theta^{(1)}|\alpha = 1) D_\theta J_i' a_i - \Phi_1^{-1}(\theta^{(1)} - \mu_1)$$

where $a_i = (a_{1i}, ..., a_{mi})'$,

$$D_\theta = \text{diag}\left\{1_1, ..., 1_m, \exp\left(\theta_{m+1}\right), ..., \exp\left(\theta_{2m}\right)\right\} \ ,$$

and

$$v_i(\theta^{(\alpha)}|\alpha) = \frac{D[P_i(\theta^{(\alpha)}|\alpha) - c_i][u_i - P_i(\theta^{(\alpha)}|\alpha)]}{(1 - c_i)\, P_i(\theta^{(\alpha)}|\alpha)} \ ,$$

and where diag denotes a diagonal matrix.

Second Derivatives

The second derivative matrix is used in the iterative procedure (13) for calculation of the mode, and also in the Taylor series expansion (12). This matrix (minus the observed information matrix) is given by

$$\frac{\partial^2}{\partial(\theta^{(0)})^2} l_0(\theta^{(0)}|u) \equiv -I_0(\theta^{(0)}|u) \equiv \sum_{i \in A} w_i(\theta^{(0)}|\alpha = 0) a_i a_i' - \Phi_0^{-1} \ ,$$

and

$$\begin{aligned} \frac{\partial^2}{\partial(\theta^{(1)})^2} l_1(\theta^{(1)}|u) \ &\equiv \ -I_1(\theta^{(1)}|u) \\ &\equiv \ \sum_{i \in A} w_i(\theta^{(1)}|\alpha = 1) D_\theta J_i' a_i a_i' J_i D_\theta + Z_i - \Phi_1^{-1} \ , \end{aligned}$$

where Z_i is a $2m \times 2m$ diagonal matrix with elements

$$Z_i = h_i v_i\,(\theta_\alpha|\alpha)\,\mathrm{diag}\{0_1, ..., 0_m, a_1 \exp{(\theta_{m+1})}, ..., a_m \exp{(\theta_{2m})}\}$$

and where

$$w_i(\theta^{(\alpha)}|\alpha) = \frac{D^2[1 - P_i(\theta^{(\alpha)}|\alpha)][P_i(\theta^{(\alpha)}|\alpha) - c_i][c_i u_i - P_i^2(\theta^{(\alpha)}|\alpha)]}{(1 - c_i)^2\, P_i^2(\theta^{(\alpha)}|\alpha)} \ . \tag{14}$$

References

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis.* New York: John Wiley & Sons.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*, 179–188.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

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

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, *49*, 359–381.

O'Hagan, A. (1994). *Kendall's advanced theory of statistics: Bayesian inference* (Vol. 2B). London: Edward Arnold.

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

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*, 331–354.

Segall, D. O., & Moreno, K. E. (1999). Development of the Computerized Adaptive Testing version of the Armed Services Vocational Aptitude Battery. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Segall, D. O., Moreno, K. E., Bloxom, B. M., & Hetter, R. D. (1997). Psychometric procedures for administering CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 131–140). Washington, DC: American Psychological Association.

Segall, D. O., Moreno, K. E., & Hetter, R. D. (1997). Item pool development and evaluation. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 117–130). Washington, DC: American Psychological Association.

Tanner, M. A. (1996). *Tools for statistical inference: Methods for the exploration of posterior distributions and likelihood functions.* New York: Springer-Verlag.

van der Linden, W. J., & Pashley, P. J. (2000). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 1–25). Boston: Kluwer.