Running head: MCAT ITEM SELECTION

Comparing the performance of five multidimensional CAT selection procedures with different stopping rules

Lihua Yao

Defense Manpower Data Center, Monterey Bay

Abstract

Through simulated data, five multidimensional CAT (MCAT) selection procedures (Yao, 2012) with varying test lengths are examined and compared using different stopping rules. Fixed item exposure rates are used for all the items, and the Priority Index (PI) method is used for the content constraints. Two stopping rules *SE* (standarded error) and *PSER* (predicted standarded error reduction) are proposed; each MCAT selection process is stopped if either the required precision has been achieved or the selected number of items has reached the maximum limit. The five procedures are: Minimum Angle (Ag, Reckase, 2009), Volume (Vm, Segall, 1996), Minimize the error variance of the linear combination (V_1, van der Linden, 1999), Minimize the error variance of the composite score with the optimized weight (V_2, Yao, 2010a), and Kullback-Leibler information (KL, Veldkamp & van der Linden, 2002). The recovery for the domain scores or content scores and their overall score, test length, and test reliability are compared across the five MCAT procedures and between the two stopping rules. It is found that the two stopping rules are implemented successfully and that KL uses the least number of items to reach the same precision level, followed by Vm; Ag uses the largest number of items. On average, to reach a precision of $SE = 0.35$, 40, 55, 63, 63, and 82 items are needed for KL, Vm, V_1, V_2, and Ag, respectively, for the *SE* stopping rule. *PSER* yields 38, 45, 53, 58, and 68 items for KL, Vm, V_1, V_2, and Ag, respectively; *PSER* yields only slightly worse results than *SE*, but with much fewer items. Overall, KL is recommended for varying-length MCAT.

Key words: BMIRT, CAT, Domain scores, MCAT, MIRT, Multidimensional Item Response Theory, Multidimensional Information, Overall scores, Multidimensional Priority Index.

Comparing the performance of five multidimensional CAT selection procedures with different stopping rules

Introduction

A CAT selection process is a cyclical procedure that is stopped by a stopping rule (Reckase, 2009; Wainer, 2000). The stopping rule can be when a specified number of test items has been administered (fixed-length), when the estimated ability has reached the desired precision level, or when a decision has been made with the desired confidence level (varying-length). It is commonly known that in the unidimensional IRT, the SEM (standard error of measurement) is larger for examinees on either end than for those in the middle, i.e., the precisions on either end are smaller than those in the middle. For a fixed-length test, the estimated precisions are different for different examinee levels. To achieve the same level of precision, some examinees may need less test items and some may need more test items. A large estimation error results in a high misclassification rate, which can be costly; achieving the required precision is more desirable. However, there are problems with using precision rules as stopping rules. On one hand, the administered test for certain examinees may be undesirably lengthy because the required precision cannot be met. On the other hand, for certain examinees, the test may be stopped too early when administering one or two more items might have improved the precision significantly. Achieving the required precision is the ultimate goal for a test; this depends on not only the quality of the item pool, but also on the CAT selection procedures. Some research has been done on using different stopping rules for the unidimensional CAT (UCAT, Dodd, Koch, & De Ayala, 1993). There are stopping rules such as the minimum standard error stopping rule, the minimum information stopping rule (Dodd, Koch, & De Ayala, 1989), and recently, the predicted standard error reduction stopping rule (Choi, Grady, & Dodd,

in press). Previous studies on MCAT item selection procedures are conducted under the fixed-length condition (Yao, 2011a, 2012; Wang & Chang, 2011). However, MCAT selection procedures with varying test length using different stopping rules deserve more research attention.

In this study, the five MCAT procedures in Yao (2011a, 2012) are applied with varying-length using two stopping rules *SE* (standard error) and *PSER* (predicted standard error reduction), which will be described later. A process is stopped if either the required precision has been achieved or the selected number of items has reached the maximum limit. The five procedures are: Minimum Angle (Ag, Reckase, 2009), Volume (Vm, Segall, 1996), Minimize the error variance of the linear combination (V_1, van der Linden, 1999), Minimize the error variance of the composite score with the optimized weight (V_2, Yao, 2010a), and Kullback-Leibler information (KL, Veldkamp & van der Linden, 2002). An item exposure rate of 0.3 is imposed for all items; this was found to produce comparable results as using higher exposure rates. The Multidimensional Priority Index (MPI) introduced in Yao (2011a) is applied for content control. The comparison is done for simulated data with the item pool coming from real live CAT data with 912 items. The following research agenda will be addressed in this study: (a) examine and compare the performance between the two stopping rules; (b) compare the performance of the five MCAT selection procedures for each of the stopping rules. The recovery for the domain scores and overall scores, item pool usage, test reliability, and test lengths are examined for each of the various simulated conditions. Please note that MIRT model is used in this study, and that each dimension represents a content or a domain; therefore, content, domain, and dimension are used interchangeably.

Multidimensional item Response Theory (MIRT) Models

Following the notation of the MIRT model in Yao & Schwartz (2006), for a dichotomously-scored item j, the probability of a correct response to item j for an examinee with the ability $\vec{\theta} = (\theta_1, \cdots, \theta_D)$ for the multidimensional three-parameter logistic (M-3PL; Reckase, 1997) model is:

$$P_{j1} = P_{j1}(\vec{\theta}) = P(x_j = 1 \mid \vec{\theta}, \vec{\beta}_j) = \beta_{3j} + \frac{1 - \beta_{3j}}{1 + e^{(-\vec{\beta}_{2j} \odot \vec{\theta}^T + \beta_{1j})}}, \tag{1}$$

where $x_j = 0$ or 1 is the response of the examinee to item j. $\vec{\beta}_{2j} = (\beta_{2j1}, \cdots, \beta_{2jD})$ is a vector of dimension D for the item discrimination parameters. β_{1j} is the intercept and $\frac{\beta_{1j}}{||\vec{\beta}_{2j}||}$ is the difficulty parameter, β_{3j} is the lower asymptote or the guessing parameter, and $\vec{\beta}_{2j} \odot \vec{\theta}_i^T = \sum_{l=1}^{D} \beta_{2jl}\theta_{il}$. The norm or multidimensional discrimination index MDISC (Reckase & McKinely, 1991) is defined as $||\vec{\beta}_{2j}|| = \sqrt{\sum_{l=1}^{D} \beta_{2jl}^2}$. The parameters for the jth item are $\vec{\beta}_j = (\vec{\beta}_{2j}, \beta_{1j}, \beta_{3j})$. Please note that only the multidimensional three-parameter logistic model is introduced here, as the items in the simulation study are all multiple choice items.

Let

$$P_j = P_j(\vec{\theta}) = P_j(X_j \mid \vec{\theta}, \vec{\beta}_j) = P_{j1}^{1_{(X_j=1)}}(1 - P_{j1})^{1_{(X_j=0)}} \tag{2}$$

for a M-3PL item, where

$$1_{(X_j=k)} = \begin{cases} 1 & \text{if } X_j = k \\ 0 & \text{otherwise} \end{cases}$$

The likelihood equation for a response to J items $\vec{X} = (X_1, \cdots, X_J)$ at a given ability $\vec{\theta}$ is

$$L(\vec{X} \mid \vec{\theta}) = P(\vec{X} \mid \vec{\theta}, \boldsymbol{\beta}) = \prod_{j=1}^{J} P_j(X_j \mid \vec{\theta}, \vec{\beta}_j). \tag{3}$$

The following statistics will be used in estimating abilities and item selection procedures.

Statistics

First Derivative. Let $P_{j1}(\vec{\theta})$ be defined as in Equation 1, indicate the probability at ability $\vec{\theta}$. For an M-3PL item j,

$$\frac{\partial P_{j1}}{\partial \vec{\theta}} = (\frac{\partial P_{j1}}{\partial \theta_1}, \cdots, \frac{\partial P_{j1}}{\partial \theta_D}) = \frac{(1 - P_{j1})}{1 + e^{-\vec{\beta}_{2j} \odot \vec{\theta}^T + \beta_{1,j}}} \vec{\beta}_{2j}. \tag{4}$$

Since

$$\frac{1}{1 + e^{-\vec{\beta}_{2j} \odot \vec{\theta}^T + \beta_{1,j}}} = \frac{P_{j1} - \beta_{3j}}{1 - \beta_{3j}}, \tag{5}$$

therefore

$$\frac{\partial \log P_j}{\partial \vec{\theta}} = (\frac{1_{(X_j=1)}}{P_{j1}} - \frac{1_{(X_j=0)}}{1 - P_{j1}}) \frac{\partial P_{j1}}{\partial \vec{\theta}} = \frac{(X_j - P_{j1})(P_{j1} - \beta_{3j})}{P_{j1}(1 - \beta_{3j})} \vec{\beta}_{2j}. \tag{6}$$

Finally,

$$\frac{\partial \log L(\vec{X} \mid \vec{\theta})}{\partial \vec{\theta}} = \sum_{j=1}^{J} \frac{\partial \log P_j}{\partial \vec{\theta}}. \tag{7}$$

Second-derivative. For an M-3PL item j,

$$\frac{\partial^2 \log P_j}{\partial \vec{\theta}^2} = \frac{\beta_{3j} X_j - P_{j1}^2}{P_{j1}^2} \frac{1}{1 - \beta_{3j}} \frac{\partial P_{j1}}{\partial \vec{\theta}} \otimes \vec{\beta}_{2j} = \frac{(1 - P_{j1})(P_{j1} - \beta_{3j})(\beta_{3j} X_j - P_{j1}^2)}{P_{j1}^2 (1 - \beta_{3j})^2} \vec{\beta}_{2j} \otimes \vec{\beta}_{2j}. \tag{8}$$

Here $\otimes$ is a vector product; $\vec{\beta}_{2j} \otimes \vec{\beta}_{2j}$ is a $D \times D$ matrix, and its mth row and nth column element is the product of the mth and nth element of $\vec{\beta}_{2j}$. Let

$$J(\vec{\theta}) = \frac{\partial^2 \log L(\vec{X} \mid \vec{\theta})}{\partial \vec{\theta}^2} = \sum_{j=1}^{J} \frac{\partial^2 \log P_j}{\partial \vec{\theta}^2}. \tag{9}$$

Test Information Function and Standard Error of Measurement (SEM)

For an item j, following M-3PL, the information function at $\vec{\theta}$ is

$$I_j(\vec{\theta}) = -E \frac{\partial^2 \log P_j}{\partial \vec{\theta}^2} = \frac{(P_{j1} - \beta_{3j})^2 (1 - P_{j1})}{P_{j1}(1 - \beta_{3j})^2} \vec{\beta}_{2j} \otimes \vec{\beta}_{2j}.$$

The test information for J items at $\vec{\theta}$ is $\mathbf{I}_J(\vec{\theta}) = \sum_{j=1}^{J} I_j(\vec{\theta})$. The composite score $\theta_{\vec{\alpha}} = \sum_{l=1}^{D} \theta_l w_l$ has a standard error of measurement $SEM(\theta_{\vec{\alpha}}) = V(\theta_{\vec{\alpha}})^{1/2}$, where $V(\theta_{\vec{\alpha}}) = \vec{w} V(\vec{\theta}) \vec{w}^T$, $\vec{w} = (w_1, \cdots, w_D) = (cos^2\alpha_1, \cdots, cos^2\alpha_D)$. Here $\cos(\vec{\alpha}) = (\cos\alpha_1, \cdots, \cos\alpha_D)$, and α_l is the angle between vector $\vec{\theta}$ and the θ_l axis for

$l = 1, \cdots, D$. $V(\vec{\theta})$ can be approximated by $I(\vec{\theta})^{-1}$. The SEM for each domain can be derived by using the angle for that dimension/domain to be 0, and all others 90^0.

Bayesian Statistics

Suppose the prior of the population is $f(\vec{\theta})$, then the posterior density function of $\vec{\theta}$ is

$$f(\vec{\theta} \mid \vec{X}) \propto L(\vec{X} \mid \vec{\theta}) f(\vec{\theta}), \tag{10}$$

and if the prior is normal $N(\mu, \Sigma)$, then

$$f(\vec{\theta}) = (2\pi)^{-D/2}(\mid \Sigma \mid)^{-1/2} exp(-\frac{1}{2}(\vec{\theta} - \vec{\mu})^T \Sigma^{-1}(\vec{\theta} - \vec{\mu})), \tag{11}$$

where $\vec{\mu}$ and Σ represent the population mean and variance-covariance matrix, respectively. Since the difference between the the right hand side and the left hand side of Equation 10 is a constant multiple, and finding the maximum value for both of them should be the same, therefore, the right hand side is used as the posterior distribution.

First-Derivative.

$$\frac{\partial \log f(\vec{\theta} \mid \vec{X})}{\partial \vec{\theta}} = \frac{\partial \log L(\vec{X} \mid \vec{\theta})}{\partial \vec{\theta}} - \frac{\partial(\vec{\theta} - \vec{\mu})}{\partial \vec{\theta}} \Sigma^{-1}(\vec{\theta} - \vec{\mu}), \tag{12}$$

where $\frac{\partial(\vec{\theta} - \vec{\mu})}{\partial \theta_k} = (0, \cdots, 1, 0, \cdots, 0)_{1 \times D}$ and 1 is in the kth position.

Second-Derivative.

$$\frac{\partial^2 \log f(\vec{\theta} \mid \vec{X})}{\partial \vec{\theta}^2} = \frac{\partial^2 \log L(\vec{X} \mid \vec{\theta})}{\partial \vec{\theta}^2} - \Sigma^{-1} = J(\vec{\theta}) - \Sigma^{-1}. \tag{13}$$

Item and test information function. Posterior test information at $\vec{\theta}$ for selected $j-1$ items is

$$\mathbf{I}_{j-1}(\vec{\theta}) = -EJ(\vec{\theta}) + \Sigma^{-1} = \sum_{m \in M-3PL} \frac{(P_{m1} - \beta_{3m})^2 (1 - P_{m1})}{P_{m1}(1 - \beta_{3m})^2} \vec{\beta}_{2m} \otimes \vec{\beta}_{2m} + \Sigma^{-1}. \tag{14}$$

Here bold variable $\mathbf{I}_{j-1}$ indicate sum of $I_1, \cdots, I_{j-1}$. The SEM for the Bayesian version is similarly derived as the nonBayesian version.

MIRT Ability Estimation Methods

For the maximum likelihood estimation methods, MIRT ability is estimated by finding the mode $\hat{\vec{\theta}}$ that maximize the likelihood function $L(\vec{X} \mid \vec{\theta})$, i.e.,

$$\frac{\partial \log L(\vec{X} \mid \vec{\theta})}{\partial \vec{\theta}} \mid \hat{\vec{\theta}} = 0 \tag{15}$$

Using Newton-Raphson method, suppose $\vec{\theta}^m$ is the m-th approximation that maximize $\log L(\vec{X} \mid \vec{\theta})$, then

$$\vec{\theta}^{m+1} = \vec{\theta}^m - \vec{\delta}^m, \tag{16}$$

where

$$\vec{\delta}^m = [\mathbf{J}(\vec{\theta}^m)]^{-1} \times \frac{\partial \log L(\vec{X} \mid \vec{\theta})}{\partial \vec{\theta}}, \tag{17}$$

and $\mathbf{J}(\vec{\theta})$ is the matrix of the second partial derivative.

In this study, MAP (Maximize a posterior) is used by finding the mode that maximizes the posterior likelihood function $f(\vec{\theta} \mid \vec{X})$. This is similar to the MLE method, but the function used is the product of the likelihood and the prior, instead of only the likelihood. Therefore, the Newton-Raphson method by Equation 16 and 17 can be applied with Bayesian version for the two terms. For the domain scores, the updated scores after each item selection and the final scores are obtained through the MIRT Bayesian maximize a posterior using standard multivariate normal as the prior. Yao (2010b) has found that MAP yields better precision than MLE and perform similarly or better than EAP (Expected a posterior). Although strong priors can be applied and may yield better precisions, they are not applied in this study and will not affect the conclusion of the study.

MCAT Item Selection Procedures

The two stopping rules are described in this section. Content constraints and item exposure control are applied in this study. The stopping rules, the content constraints, and

the item exposure control are combined to form the probability or weight in item selection.

The Priority Index for Content Constraints

The multidimensional priority index (MPI, Cheng & Chang, 2009; Yao, 2011a) for each item j is defined by

$$MPI_j = \prod_{l=1}^{D} f_{jl}^{c_{jl}}, \tag{18}$$

where the constraint matrix $C = (c_{jl})_{J \times D}$, indicating the loading information for item j on domain l, is defined as the following:

$$c_{jl} = \begin{cases} 1 & \text{if item j load on domain l} \\ 0 & \text{otherwise} \end{cases}$$

If the percentage of selected items is fixed for each domain, then f_{jl} is defined by the difference of the number of requested items T_l and the number of selected items t_l for domain l divided by T_l. At the beginning, the ratio is 1 when no items have been selected from domain l. The ratio shrinks as items from this domain are selected. When the number of selected items reaches the required number of items for domain l, the ratio f_l is 0; no more items will be selected from this domain.

The percentage of selected items for each domain can be flexible. Let the lower and upper bound in selecting items for domain l be L_l and U_l. Let $L = \sum_{l=1}^{D} L_l$. In selecting nth items, with $n < L$, define

$$f_{jl} = \max\{\frac{L_l - t_l}{L_l}, 0\}. \tag{19}$$

When $n \geq L$, define

$$f_{jl} = \max\{\frac{U_l - t_l}{U_l}, 0\}. \tag{20}$$

Item Exposure Control

For the jth item, let r_j denote its exposure rate. For each selection step, let n_j be the number of examinees that have already selected item j. The index for the item

exposure control is defined by (Chang, in press; Linden & Veldkamp, 2004, 2007; Yao, 2011a)

$$f_{jl} = \max\{\frac{r_j - \frac{n_j}{N}}{r_j}, 0\}, \tag{21}$$

where N is the total number of examinees. This index will make sure that no item is selected with exposure rate larger than the predefined rate $\vec{r} = (r_1, \cdots, r_J)$.

The SE and PSER Stopping Rules

Let $\vec{P} = (p_1, \cdots, p_D)$ represents the required SEM for the D domain ability estimates; the smaller the SEM, the larger the precision. Let $\hat{P} = (\hat{p}_1, \cdots, \hat{p}_D)$ be the SEM estimates based on the current selected items. Combining the item exposure rate, content constraints with upper and lower limit for each domain, and the estimated domain score precision, define the index in Equation 18 to be the following for the SE stopping rule.

$$f_{jl} = [\max\{[1-(\frac{p_l}{\hat{p}_l})^a+\epsilon_1], 0\}][\max\{(\frac{r_j - \frac{n_j}{N}}{r_j}), 0\}][(1_{t_l \le L_l}(\frac{L_l - t_l}{L_l}+\epsilon_2/t_l)+1_{t_l > L_l}\max\{1-(\frac{t_l}{U_l})^b, 0\})]. \tag{22}$$

If for some domain l, the precision has been achieved, then the items loading in domain l will not be selected anymore (the first term in Equation 22). If an item has been selected more times and has reached the required exposure rate, then it will be not be selected anymore (the second term in Equation 22). If the number of selected items has reached the maximum limit for certain domain, then no more items will be selected from that domain (the third term in Equation 22); the first part of the third term is to ensure that each domain will have the minimum required number of items. For $f(x) = 1 - x^n$, where $0 \le x \le 1$, the power n decide the shape of the curve. For precision, the smaller the power a, the larger the weight that is put on the precision. Suppose there are two contents, with one having a large SEM and one having a small SEM; then $\frac{p_l}{\hat{p}_l}$ will be small and large for the two contents, respectively. Suppose they are 0.3 and 0.7 for the two contents

respectively. $1-0.3^2=0.91$, $1-0.3^3=0.973$, $1-0.7^2=0.51$, $1-0.7^3=0.657$, and $0.91/0.51=1.78>0.973/0.657=1.48$. With the other terms in Equation 22 remaining the same, it will have a higher probability of selecting items in the first content and a much lower probability of selecting items in the second content when using $a=2$ rather than $a=3$. In general, since $f^{'}(x)=-nx^{n-1}$, the rate of change is even larger when $n<1$. That is to say, more weight will be imposed for the precision when the power a is small. The items in the content with better precision will have a much lower probability of being selected when the power a is small. At the beginning of a selection process, $p_l<\hat{p}_l$, the selection process stops for that domain if $p_l\geq\hat{p}_l$. Here ϵ_1 is a small number that can be adjusted so that the precision of the estimates can be slightly above the required precision. ϵ_2 is a small number that can be adjusted so that the minimum required number of items for each domain can be administered first. Similarly, the smaller the power b, the larger the weight that is put on the length restriction condition. In this study, the precision index is applied after the minimum number of items has been administered, and $a=2$, $b=3$, $\epsilon_1=0.001$, and $\epsilon_2=1$. $a<b$ was chosen for the purpose of putting more weight on the precision than on the maximum required item number. Different values for a and b can be applied and tested for the purpose of giving more or less weight to the precision and to the maximum required item number, respectively.

Pre-runs for some procedures showed that for some examinees, the administered test is lengthy, with too many items being administered without an accompanying improvement in precision. Therefore, a modified procedure (PSER) that predicts the reduction of the SE is proposed and compared with the SE method. For PSER, there are two modifications: 1) a predetermined parameter $\alpha=0.03$ is applied, and if the SEM reduction based on the current selected items and the previously selected items is smaller than α, then the item selection for this domain is stopped, even if the SE requirement has not been met; 2) a predetermined parameter $\beta=0.05$ is applied, and if the SE reduction

based on the current selected items and the previously selected items is larger than β, then the item selection for this domain will continue with a slightly larger weight (adding .0001), even if the SE requirement has been met. For unidimensional IRT, the information is a monotonically increasing function with respect to the number of items administered. However, this is not the case with the MIRT models. At each score point vector, the information is a matrix and the directional information along each of the dimensions/domains is not a monotonic function with respect to the number of items administered. Therefore, extra rules are applied. They are: 3) if the current precision is not smaller than the previous step and the current precision is within α distance away from the required precision, then stop selecting items from this domain; 4) if the current precision (SEM) is not smaller than the previous step and the current precision is outside $2 \times \beta$ distance away from the required precision, then the item selection for this domain will continue with slightly higher weight (adding .0001). Please note that α and β can be specified differently and $\beta > \alpha$. β measures how much SEM reduction you would allow to select one more item to increase the precision. α measures how much SEM reduction that you can tolerate to keep selecting items. The rules for PSER ensure that (a) lengthy tests are prevented when the pool has no more quality items that will improve the precision of the examinee's ability estimates; (b) better precision is obtained with one or two more items; and (c) the precision is not much worse than the required precision.

Please note those SEMs or the precisions for the domain abilities are implemented in the probability in selecting items as presented in Equation 22, other criteria are possible and deserve study. For example, one can modify Equation 22 by adding the SEM for the overall ability, or one can delete the SEMs for the domain ability and use only the SEM for the overall ability. The choice depends on the purpose of the test.

The five item selection procedures described in Yao (2012) are briefly described in this section. For all the procedures, the initial abilities are set to $\theta_l = 0$ for $l = 1, \cdots, D$.

For $j = 2, \cdots, J$, suppose $j-1$ items have been selected. To select the next jth item, suppose the updated ability is $\vec{\theta}^{j-1}$. For each of the procedures, the steps proposed are repeated. Please note that the procedures are for both Bayesian and non-Bayesian; for Bayesian, add Σ^{-1} to the information.

Select an item that has the maximum information in the direction that has the minimum information for previously selected items–*Angle* **or** *Ag*

1. At the ability level $\vec{\theta}^{j-1}$, let the direction $\vec{\alpha} = (\alpha_1, \cdots, \alpha_D)$ be the minimizer such that $\cos(\vec{\alpha})\mathbf{I}_{j-1}(\vec{\theta}^{j-1})\cos(\vec{\alpha})^T$ has a minimum value for all possible angles. Here $\cos(\vec{\alpha}) = (\cos\alpha_1, \cdots, \cos\alpha_D)$.

2. For each item m in the pool, compute

$$\mathbf{I}_j^m(\vec{\theta}^{j-1}) = \mathbf{I}_{j-1}(\vec{\theta}^{j-1}) + \frac{(P_{m1} - \beta_{3m})^2(1 - P_{m1})}{P_{m1}(1 - \beta_{mj})^2}\vec{\beta}_{2m} \otimes \vec{\beta}_{2m}$$

at ability $\vec{\theta}^{j-1}$.

3. Select item $j = m$ such that $\cos(\vec{\alpha})\mathbf{I}_j^m(\vec{\theta}^{j-1})\cos(\vec{\alpha})^T$ has a maximum value (among all the items in the pool).

4. Update ability $\vec{\theta}^{j}$ and information $\mathbf{I}_j(\vec{\theta}^{j})$ based on the selected j items.

Select an item that has the maximum volume or the maximum determinant of the information–*Volume* **or** Vm

In Segall (1996), he proposed selecting the next item j by maximizing the determinant of the posterior information as follows:

$$W = |\mathbf{I}_{j-1}(\vec{\theta}^{j-1}) + I_j(\vec{\theta}^{j-1}) + \Sigma^{-1}|, \tag{23}$$

where $\mathbf{I}_{j-1}(\vec{\theta}^{j-1})$ is the information obtained from already selected $j-1$ items at the ability estimates $\vec{\theta^{j-1}}$.

1. For each item m in the pool, compute the volume or the determinant of the information using

$$W_m = |\mathbf{I}_{j-1}(\vec{\theta}^{j-1}) + \frac{(P_{m1} - \beta_{3m})^2(1 - P_{m1})}{P_{m1}(1 - \beta_{mj})^2}\vec{\beta}_{2m} \otimes \vec{\beta}_{2m} + \Sigma^{-1}|$$

at ability $\vec{\theta}^{j-1}$.

2. Select item $j = m$ such that W_m has the maximum value.

3. Update ability $\vec{\theta}^{j}$ and information $I_j(\vec{\theta}^{j})$ based on the selected j items.

For the non-Bayesian procedure, the above equations still hold with the removal of Σ^{-1}. However, the first D items must be selected from the D domains, especially for the items of simple structure, as the matrix needs to be non-singular.

Select an item that has the minimum error variance for the composite score of equal weight–V_1

This method was studied in van der Linden (1999) for increasing the precision for overall scores. It is similar to V_2 described below, with the weight $\vec{w}_{j-1}$ being pre-fixed with equal values for all $j = 1, \cdots, J$, i.e., $\vec{w}_{j-1} = (w_1, \cdots, w_D), w_l = 1/D$ for $l = 1, \cdots, D$.

Select an item that has the minimum error variance for the composite score of optimized weight–V_2

For a test with J items of known item parameters, for a given score point $\vec{\theta}$, the test information is $\mathbf{I}_J(\vec{\theta})$. The composite score $\theta_{\vec{\alpha}} = \sum_{l=1}^{D} \theta_l w_l$ has a standard error of measurement $SEM(\theta_{\vec{\alpha}}) = V(\theta_{\vec{\alpha}})^{1/2}$. The weight $\vec{w}$ - the optimized weight such that $SEM(\theta_{\vec{\alpha}})$ has a minimum value - does exist (Yao, 2010a, 2012). The weight for selecting j items $\vec{w}_j = \vec{w}$ is the optimized weight derived on the estimated domain abilities and the elected items.

The following steps are used in selecting items for V_2. Let $M < J$ be a chosen integer.

1. For $j <= M$, the weight is pre-fixed weight of equal values, i.e., $\vec{w}_{j-1} = (w_1, \cdots, w_D), w_l = 1/D$ for $l = 1, \cdots, D$.

2. For $j > M$, compute the optimized weight $\vec{w}_{j-1}$ based on the $j-1$ selected items.

3. Select item $j = m$ such that $\vec{w}_{j-1}[\mathbf{I}^m_{j-1}(\vec{\theta}^{j-1})]^{-1}(\vec{w}_{j-1})^T$ has a minimum value.

4. Update ability $\vec{\theta}^j$ and information $\mathbf{I}_j(\vec{\theta}^j)$ based on the selected j items.

The integer M can be chosen by the user. For example, $M = 0$ or $M = \frac{J}{3}$, where J is the total number of selected items. $M = 0$ is applied in this study; pre-run shows that results from $M = 0$ and $M = \frac{J}{3}$ are similar.

Select an item that has the maximum posterior KL information–KL

For a M-3PL item m, the Kullback-Leibler information is the distance between two likelihoods at two ability points $\vec{\theta}^{j-1} = (\theta_1^{j-1}, \cdots, \theta_D^{j-1})$ and $\vec{\theta}_0$ and is defined as:

$$K_m(\vec{\theta}^{j-1}, \vec{\theta}_0) = E_{\vec{\theta}_0} \log[\frac{P_m(X_m \mid \vec{\theta}_0, \vec{\beta}_m)}{P_m(X_m \mid \vec{\theta}^{j-1}, \vec{\beta}_m)}] = P_{m1}(\vec{\theta_0}) \log \frac{P_{m1}(\vec{\theta}_0)}{P_{m1}(\vec{\theta}^{j-1})} + (1 - P_{m1}(\vec{\theta}_0)) \log \frac{1 - P_{m1}(\vec{\theta}_0)}{1 - P_{m1}(\vec{\theta}^{j-1})}, \tag{24}$$

where $\vec{\theta}_0$ is the true ability, and $\vec{\theta}^{j-1}$ is the current ability estimates based on selected $j-1$ items. The Kullback–Leibler information tells us how well the response variable discriminates between the ability estimates and the true ability value. For $j-1$ selected items, define $\mathbf{K_{j-1}}(\vec{\theta}^{j-1}, \vec{\theta}_0) = \sum_{l=1}^{j-1} K_l(\vec{\theta}^{j-1}, \vec{\theta}_0)$. The Bayesian KL for item m (Chang & Ying, 1996; Veldkamp & van der Linden, 2002) is

$$\begin{aligned} K_m(\vec{\theta}^{j-1} \mid \vec{X}) &= \int_{\vec{\theta}} (\mathbf{K}_{j-1}(\vec{\theta}^{j-1}, \vec{\theta}) + K_m(\vec{\theta}^{j-1}, \vec{\theta})) f(\vec{\theta} \mid \vec{X}) d\vec{\theta} \\ &= \int_{\theta_1^{j-1} - \delta_j}^{\theta_1^{j-1} + \delta_j} \cdots \int_{\theta_D^{j-1} - \delta_j}^{\theta_D^{j-1} + \delta_j} (\mathbf{K}_{j-1}(\vec{\theta}^{j-1}, \vec{\theta}) + K_m(\vec{\theta}^{j-1}, \vec{\theta})) f(\vec{\theta} \mid \vec{X}) d\theta_1 \cdots \theta_D \end{aligned} \tag{25}$$

where $\delta_j = \frac{3}{\sqrt{j}}$.

1. For each item m in the pool, compute the posterior KL information $K_m(\vec{\theta}^{j-1} \mid \vec{X})$ using Equation 25. Here $\vec{X}$ is the response vector for the selected $j-1$ items.

2. Select item $j = m$ such that $K_m(\vec{\theta}^{j-1} \mid \vec{X})$ has the maximum value.

3. Update ability $\vec{\theta}^{j}$ based on the selected j items.

For the five selection procedures, each of the selection criteria is modified by multiplying the MPI defined in Equation 18 and 22. For example, for the Ag method, step 3 is modified to be: select item $j = m$ such that $\cos(\vec{\alpha})\mathbf{I}_j^m(\vec{\theta}^{j-1})\cos(\vec{\alpha})^T MPI_m$ has a maximum value (among all the items in the pool). Similarly, for Vm, V_1 and V_2, and KL, step 2, 3, and 2 are modified by multiplying MPI_m, respectively, to each of the selection criteria in selecting the next items.

Applications

Item Pool

The item parameters in the item pool come from the item parameter estimates of live data for approximately 176,000 examinees taking the CAT Armed Services Vocational Aptitude Battery (ASVAB) to qualify for service in the U.S. Military. Each examinee took four tests of 55 items: 15 Arithmetic Reasoning (AR), 15 Word Knowledge (WK), 10 Paragraph Comprehension (PC), and 15 Math Knowledge (MK). For the 176000 examinees, with 49000 responses for each item, there are 257 AR items, 265 WK items, 144 PC items, and 246 WK items, with a total of 912 items. The MIRT four-dimensional confirmatory analysis following M-3PL model with simple structure is conducted, with the scale fixed by imposing the prior of the ability distribution to be multivariate normal with a mean of (0,0,0,0) and a variance-covariance matrix of $\mathbf{A}$ obtained from the total row

score of the responses, as shown below.

$$\mathbf{A} = \begin{pmatrix} 1 & 0.5 & 0.5 & 0.7 \\ 0.5 & 1 & 0.6 & 0.4 \\ 0.5 & 0.6 & 1 & 0.4 \\ 0.7 & 0.4 & 0.4 & 1 \end{pmatrix}_{4\times 4} . \tag{26}$$

The item summary statistics for the item pool are displayed in Table 1.

Simulation Conditions

Sample size of 1000 examinees are simulated from a multivariate normal distribution with a mean of (0,0,0,0) and correlations between the domains are .6. For each of the four "true" domain abilities AR, WK, PC, and MK, the "true" overall ability $AFQT$ (Armed Forces Qualification Test) can be defined by the simple averaging of the four domain scores or by finding the optimized weight for the given score points based on the selected item parameters. For each of the 1000 examinees and each selection procedure, 10 different seeds are used, resulting in 10 replications.

For each of the five procedures, the MPI index defined in Equation 18 and 22 is implemented. The item exposure rate is defined to be $r_j = 0.3$ for all items in the pool. Three levels of precision (SEM) are set to be: $\vec{P}_1 = (.35, .35, .35, .35)$, $\vec{P}_2 = (.3, .3, .3, .3)$, and $\vec{P}_3 = (.25, .25, .25, .25)$, indicated by condition $R1 - R3$, respectively. The level of precision tells us how certain our estimation is; it gives us a confidence interval about where the true values might be. The smaller the level is, the more accurate that we can say about our estimate. The choice of those numbers and the choice of α and β depends on the purpose of the test and the quality of the item bank and should be considered carefully; smaller number for higher precision is desirable, but if the the items in the item pool cannot give such precision, the test would be lengthy. The minimum and maximum number of items required for each of the four domains is set to be 5 and 30. For the three

precision levels, PSER is applied and the conditions are indicated by $R4 - R6$ accordingly. In summary, the varying conditions are : (a) one set of populations of size 1000; (b) ten replications; (c) five selection procedures; (d) two stopping rules; and (e) three precision levels.

MIRT Domain Ability and Overall Ability Estimates

For V_1, the "true" and the final estimated overall scores are a simple average of the four domain scores. For Ag, Vm, V_2, and KL, the "true" and the final overall scores are obtained from the weighted sum of the "true" and estimated four domain scores, respectively, with the optimized weight at the estimated domain score points; the optimized weights are updated after each selected item and the final optimized weights for the overall score are obtained based on the selected item parameters and the updated domain score estimates.

Evaluation Criteria

For each of the 1000 examinees, the $ABSBIAS$ and $BIAS$ for the ten replications are derived, for the purposes of examining the stability in item selection and estimating the abilities. For each condition and each examinee $i = 1, \cdots, N$, define $BIAS_i = \frac{\sum_{p=1}^{n}(g_{ip} - g_{itrue})}{n} = g_i - g_{itrue}$ and $ABSBIAS_i = \frac{\sum_{p=1}^{n}|g_{ip} - g_{itrue}|}{n}$, where n indicates the number of replications, and g_{ip} and g_{itrue} indicate the score estimates for the pth replication and the true values. The final estimates for the ith examinee is $g_i = \frac{1}{n}\sum_{p=1}^{n} g_{ip}$. Here g is a general term that may represent domain scores or overall scores. The *correlation* between the estimates and the true is defined by $\frac{1}{n}\sum_{p=1}^{n} corr(g_p, g_{true})$, with $g_p = (g_{1p}, \cdots, g_{Np})$ and $g_{true} = (g_{1true}, \cdots, g_{Ntrue})$. For each $MCAT$ item selection procedure and each population, the absolute bias $ABSBIAS$ for the four domain scores and the composite score $AFQT$ are computed by averaging over those

for the $N = 1000$ simulated examinees and is defined by
$ABSBIAS = \frac{1}{N}\sum_{i=1}^{N} ABSBIAS_i$. The *MCAT* item selection procedures are compared by examining *test reliability* (averaging over the $n = 10$ replications), *ABSBIAS*, *BIAS*, and the *correlation* between the estimates and the true for the four domains and the overall scores.

The *test reliability* for the UCAT was introduced in Segall (2001) and was extended to MCAT in Yao (2012); the computation used 7^4 $\vec{\theta}$ point and each with 500 simulees. It will be examined in this study; the computation time for the test reliability is lengthy.

Results

High quality items tend to be administered for the examinees taking the test earlier, resulting in a lower probability of being chosen for later examinees. This may create a problem where the ability estimates for those examinees taking the test earlier have more precision than the ability estimates for those examinees taking the test at a later time, as high quality items have more information than low quality items. Different methods can be applied to fix or mitigate this problem. One way to fix this problem is by using multi-stage Alpha Stratification (Chang & Ying, 1999; van der Linden, 2005) method, administering "low information" items first. Another way is to use random selection method from candidate pools, similar to shadow test. The following modification is applied in this study:

At the beginning of each selection process, the 1000 randomly simulated examinees are assigned to an initial ability value of 0. They then "took" the test in the order that was simulated. For the first few selected items, the ability estimates are not accurate. Therefore, to balance the item usage and to control the exposure rate, the following rule is applied: For all five selection procedures, the first item is randomly chosen from the top M_1 best items based on the selection criteria at ability level $(0, 0, 0, 0)$. The next three

items are randomly chosen from the top $M_2 - 2$, $M_3 = M_2 - 4$, and $M_4 = M_2 - 6$ items, respectively, at the current updated ability level. The rest of the selection follows the procedures described earlier. In running the program, the M_1 and M_2 are specified by the user. If M_1 and M_2 are small, then the order of the examinees towards the end might need more items than those at the beginning in order to achieve the same precision, as the item exposure control rate is enforced and some "good" items may have been administered too many times. In this study, $M_1 = 200$, $M_2 = 100$. When $M_1 = 1$, $M_2 - 2 = 1, M_3 = 1$, and $M_4 = 1$, no randomness is applied.

To examine the effect of the examinee's order in taking the test, the average test lengths for the first 50, 100, 200, 500, 700, and 1000 examinees are produced and plotted in Figure 1. One can see that KL has the smallest test length and the examinee's order in taking the test has little effect for KL. Compared to the other procedures, Ag has the largest test length by a significant margin. Vm has the second shortest test length. As the examinee's order goes up from 100 to 200, test length for Vm shows small differences; however, as the order goes up from 200 to 500, test length for Vm increases significantly. Examinee's order in taking the test has a large impact to V_1 and V_2. As the examinee's order goes up from 50 to 500 to 1000, test length for V_1 and V_2 shows larger changes. These phenomena are consistent with previous research conclusions (Yao, 2011a): KL tends to select a narrow range of items, followed by Vm. V_1 and V_2 tend to select a wide range of items. Figure 1 also shows that *PSER* has smaller test length than *SE* ($R4$ versus $R1$, $R5$ versus $R2$, and $R6$ versus $R3$).

Recovery of the Domain Abilities and Overall Abilities

The four domain ability estimates and their overall ability estimates are compared with their "true" values by examining their *correlation*, *BIAS*, and *ABSBIAS*.

Figure 2 has two plots representing the *correlation* and *ABSBIAS*. The x-axis on

each plot represents the five procedures with three precision levels, indicated by $R1 - R3$ for *SE*, and by $R4 - R6$ for *PSER*. The y-axis represents the *correlation* or *ABSBIAS* for AR, WK, PC, MK, and AFQT, respectively. For each MCAT item selection method, as the required *SEM* rate p goes from .35 to .2 for conditions $R1$ to $R3$, the *correlation* becomes larger and the *ABSBIAS* becomes smaller. Figure 2 also shows that *PSER* has smaller *correlation* and larger *ABSBIAS* than those for *SE* with the same precision levels. For each condition, the precision for the five MCAT methods are similar, as they should be, since the precision is the stopping rule.

To examine the ability recovery in detail, *BIAS*'es for the 1000 examinees for the five procedures for condition $R4$ are plotted against their true values, as shown in Figure 3. The first two rows have the *BIAS*'es for the four domain scores for KL. It is clear that the four domains have similar recovery, as expected. The next two rows compare the *BIAS*'es for the domain AR for the other four selection methods. The results are similar to each other as well as to the results from method KL. The remaining five plots compare the recovery for the five procedures for the overall score $AFQT$. The five methods perform similarly.

Test Length and Item Pool Usage

For each varying condition and each simulee, the *test length* and *SEM* are the average of the 10 replications. For each domain, each simulee is classified into 6 levels: $\{\theta < -2\}$, $\{\theta \geq -2\} \cap \{\theta < -1\}$, $\{\theta \geq -1\} \cap \{\theta < 0\}$, $\{\theta \geq 0\} \cap \{\theta < 1\}$, $\{\theta \geq 1\} \cap \{\theta < 2\}$, and $\{\theta \geq 2\}$. For each of the four domains, there are approximately 26-27, 120-150, 310-360, 120-130, and 20-30 simulees classified into levels 1-6, respectively; for a normally distributed population $N(0, 1)$, there is about 6% of populations fall outside the range of $(-2, 2)$. For demonstration purposes, the results of classifying simulees into 6 levels by domain AR are displayed .

For each classified level, the mean of the standard error of measurement and the mean of the *test length* for those classified into the 6 levels are derived and presented in Table 2 for conditions $R3$ and $R6$. The required precisions are .25 and the stopping rules are SE and $PSER$, respectively. For Ag, it is observed that the *test length* is long and the precision is not good relative to the other four methods. Examining the precision, one can see that the process SE and $PSER$ have been implemented successfully, with the precision meeting the requirement. For level $R3$, all the precisions are smaller than 0.25, except in the cases where the maximum required item number is met. For example, for method KL and examinees in level 6, AR and PC have precision .27, larger than 0.25, but the items selected in both AR and PC have reached the maximum number of 30. Examinees in level 6 were administered more items than examinees in the other levels. For $R6$ ($PSER$), the precisions are around .25, but the test lengths are much smaller than those for $R3$ (SE).

For all conditions $R1 - R6$, the item pool usage is 100 percent for all methods except KL, with a rate of 77%, 82%, 88%, 73%, 81%, and 85% for $R1$-$R6$, respectively; for KL, $PSER$ has slightly smaller item usage than the corresponding SE. This is consistent with the observation that $PSER$ has smaller test length than the corresponding SE. For domain PC, the item pool usage is 100% for all five methods. Examining the precision, one also observes that PC has worse precision than the other domains and that the maximum required number of item for PC has been reached in many cases. In Table 1, it is clear that PC has smaller discrimination than the other domains. Therefore, items in PC have smaller information than the others. We can conclude that varying-length tests might be a problem for a domain for an item pool with less information than the others. If one needs to use varying-length test method, less precision or smaller maximum number of items should be specified for this domain.

The distributions of test length for the five methods for condition $R1$ are plotted in Figure 4; they all resemble a normal distribution. It is clear that KL has the shortest test

length and a narrow distribution (smaller deviation). The distribution for Vm is wider than KL (larger deviation); for Vm, there are short length tests and there are long length tests, depending on the examinees. V_2 has a narrow distribution than V_1; overall test length for V_2 is smaller than the overall test length for V_1. Ag has a narrow distribution and the peak occurred at a place that is higher than the other four procedures; many examinees have a long test using Ag.

Discussion

In Yao (2011a), five MCAT item selection procedures with fixed-length with MPI for content constraints and item exposure control are examined through simulated data. It is found that imposing an exposure limit of .3 yields comparable results as imposing a higher exposure rate. Although the Sympson-Hetter procedure (1985) yields a slightly higher precision, it is not applied here, as it has a lower item pool usage. Therefore, an exposure limit of .3 is applied in this study. Instead of fixed-length selection, this study employs varying-length as the stopping rule, with three levels of precision as the varying conditions. Two stopping rules *SE* and *PSER* are proposed. Through combining the stopping rules, content constraints, and item exposure rate, the probability or weight in selecting items is applied and found to be successful. The defined probability index can be varied by varying the weight for the precision and the content constraints. It is found that *PSER* yields less precision but has a smaller test length than *SE*. For the five selection methods, KL tends to select a narrow range of items. KL has the smallest test length, followed by Vm, V_1, and V_2. Ag has the longest test length. For Ag, items are selected in the direction that has minimum information. Items in domains with less information (for example, *PC*) tend to be selected first. Such domains tend to reach the maximum required item number before the precision has been reached for varying-length method. When deciding which method to use (fixed-length versus varying-length), precision level,

and the maximum required item number, item statistics for the item pool are an important consideration factor. Varying-length method shortens the test length if the item pool has enough "high quality" items; those "high quality" items should be spread across different ability levels. Studies varying those factors and comparing the performance of fixed-length versus varying-length are needed. Varying-length tests can still be used, but may end up reaching the maximum required test length for many examinees if the item pool does not contain enough "high quality " items.

High quality items tend to be administered to examinees who take the test earlier. In this study, the first few items are randomly chosen. It is found that for KL, the examinee's order in taking the test has a small effect on its test length. The order affects V_1 and V_2 the most. For Vm, the examinee's order for the first 200 does not affect its test length as much as it does for later orders; the random number chosen is M_1=200. Larger numbers for M_1 are desirable, especially for Vm, V_1, and V_2. Other methods, such as multi-stage Alpha Stratification, should be implemented in the future. Overall, KL is recommended for varying-length MCAT.

In this study, the item pool contains items of simple structure. The computer programs, SimuMCAT (Yao, 2011b), for the five procedures can be applied for pools containing complex structured items; they are available for free download at www.BMIRT.com.

Acknowledgments

I would like to thank the reviewers and the editor for their valuable input on the earlier version of this manuscript. I would also like to thank my daughter Sophie Chen for her editorial assistance. The views expressed are those of the author and not necessarily those of the Department of Defense or the United States government.

References

Chang, H. (2011, in press). *Making computerized adaptive testing diagnostic tools for schools.* In R. W. Lissitz & H. Jiao (Ed.), Computers and their impact on state assessment: Recent history and predictions for the future. Information Age Publisher.

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, *20*, 213–229.

Chang, H., & Ying, Z. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, *23(3)*, 211-222.

Cheng, Y. & Chang, H. H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology, 62*, 369-383.

Choi, S. W., Grady, M., & Dodd, B. G. (in press). A new stopping rule for computerized adaptive testing. *Educational and Psychological Measurement.*

Dodd, B. G., Koch, W. R., & De Ayala, R. J. (1989). Operational characteristics of adaptive testing procedures using the graded response model. *Applied Psychological Measurement, 13*, 129-143.

Dodd, B. G., Koch, W. R., & De Ayala, R. J. (1993). Computerized adaptive testing using the partial credit model: Effects of item pool characteristics and different stopping rules. *Educational and Psychological Measurement*, *53*, 61-77.

Mulder, J., & van der Linden, W.J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika, 74(2)*, 273-296.

Reckase, M. D. (2009). *Multidimensional Item Response Theory.* New York: Springer.

Reckase, M.D., & McKinely, R.L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement*, *15*, 361-373.

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

Segall, D. O. (2001). General ability measurement: An application of multidimensional item response theory. *Psychometrika*, *66*, 79-97.

van der Linden, W. J. (1999). Multidimensional adaptive testing with a minimum error-variance criterion. *Journal of Educational and Behavioral Statistics*, *24*, 398–412.

van der Linden, W.J. (2005). Linear models for optimal test design. New York: Springer. *http://www.springer.com/statistics/social/book/978-0-387-20272-3.*

van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics, 29,* 273-291.

van der Linden, W. J., & Veldkamp, B. P. (2007). Conditional item-exposure control in adaptive testing using item-ineligibility probabilities. *Journal of Educational and Behavioral Statistics, 32,* 398-417.

Veldkamp. B.P. & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, *67*, 575-588.

Yao, L. (2003). *BMIRT: Bayesian multivariate item response theory.* [Computer software]. Monterey, CA, Defense Manpower Data Center.

Yao, L. (2010a). Reporting valid and reliability overall score and domain scores. *Journal of Educational Measurement. 47.* 339-360.

Yao, L. (2010b). *Multidimensional ability estimation: Bayesian or non-Bayesian.* Unpublished manuscript.

Yao, L. (2011a, October). *Multidimensional CAT item selection procedures with item exposure control and content constraints.* Paper presented at the 2011 International Association of Computer Adaptive Testing (IACAT) Conference, Pacific Grove, California.

Yao, L. (2011b). simuMCAT: simulation of multidimensional computer adaptive testing [Computer software]. Monterey: Defense Manpower Data Center.

Yao, L. (2012). Multidimensional CAT item selection methods for domain scores and composite scores: Theory and Applications. *Psychometrika, 77(3)*, 495-523.

Yao, L., & Schwarz, R. D. (2006). A multidimensional partial credit model with associated item and test statistics: an application to mixed-format tests. *Applied Psychological Measurement. 30*, 469-492.

Wainer, H. (ED). (2000). *Computerized adaptive testing: A primer (2nd ed.).* Hillsdale, NJ: Lawrence Erlbaum.

Wang, C., Chang, H.-H., & Boughton, K. (2011). Kullback-Leibler information and its applications in multi-dimensional adaptive testing. *Psychometrika, 76*, 13-39.

Wang, C., Chang, H.-H., & Huebner, A. (2011). Restrictive Stochastic Item Selection Methods in Cognitive Diagnostic Computerized Adaptive Testing. *Journal of Educational Measurement*, *48*. 255-273.

Table 1

Item Statistics for the Item Pool

Type	Discrimination				Difficulty				Guessing			
	AR	*WK*	*PC*	*MK*	*AR*	*WK*	*PC*	*MK*	*AR*	*WK*	*PC*	*MK*
MIN	0.63	0.27	0.83	0.16	-7.35	-8.39	-7.23	-7.92	0.05	0.08	0.08	0.062
MAX	5.84	3.96	4.78	5.99	5.55	6.37	4.58	8.99	0.35	0.29	0.22	0.392
MEAN	2.00	2.02	1.89	2.24	-0.76	-0.35	-1.40	-0.17	0.19	0.19	0.18	0.190
STD	0.59	0.59	0.54	0.86	2.55	3.07	2.58	2.72	0.03	0.02	0.02	0.038

Table 2

Precision and test length for condition R3 and R6

Method	*Level*	R3					R6				
Method	*Level*	Precision				*Test length*	Precision				*Test length*
		AR	*WK*	*PC*	*MK*		*AR*	*WK*	*PC*	*MK*	
Ag	1	0.35	0.33	0.31	0.30	120	0.35	0.33	0.32	0.32	114
Ag	2	0.31	0.32	0.31	0.29	119	0.31	0.32	0.32	0.30	111
Ag	3	0.30	0.31	0.32	0.28	119	0.30	0.32	0.32	0.29	108
Ag	4	0.29	0.31	0.33	0.28	118	0.29	0.32	0.33	0.29	108
Ag	5	0.33	0.31	0.34	0.28	118	0.33	0.32	0.34	0.30	112
Ag	6	0.39	0.31	0.36	0.30	119	0.39	0.33	0.37	0.31	115
Vm	1	0.25	0.26	0.26	0.25	95	0.27	0.28	0.28	0.26	76
Vm	2	0.24	0.26	0.26	0.24	94	0.27	0.28	0.29	0.26	72
Vm	3	0.25	0.27	0.27	0.24	102	0.27	0.28	0.30	0.26	77
Vm	4	0.25	0.27	0.29	0.24	103	0.27	0.28	0.31	0.26	80
Vm	5	0.25	0.26	0.29	0.24	100	0.27	0.27	0.31	0.26	79
Vm	6	0.29	0.24	0.30	0.24	98	0.30	0.26	0.31	0.26	81
V1	1	0.24	0.26	0.26	0.25	96	0.27	0.28	0.28	0.26	77
V1	2	0.24	0.27	0.27	0.25	97	0.27	0.28	0.29	0.26	75
V1	3	0.26	0.28	0.28	0.25	106	0.28	0.29	0.29	0.26	82
V1	4	0.26	0.28	0.29	0.25	107	0.28	0.29	0.30	0.27	84
V1	5	0.25	0.27	0.31	0.25	104	0.28	0.28	0.31	0.26	82
V1	6	0.30	0.25	0.32	0.25	101	0.30	0.27	0.32	0.27	86
V2	1	0.25	0.28	0.26	0.25	94	0.27	0.29	0.29	0.26	77
V2	2	0.25	0.29	0.28	0.26	96	0.26	0.29	0.29	0.26	77
V2	3	0.28	0.31	0.30	0.27	101	0.28	0.30	0.31	0.27	84
V2	4	0.28	0.31	0.33	0.26	101	0.28	0.30	0.32	0.27	85
V2	5	0.29	0.29	0.35	0.25	100	0.29	0.29	0.34	0.26	85
V2	6	0.34	0.26	0.36	0.25	100	0.32	0.27	0.36	0.26	88
KL	1	0.25	0.25	0.25	0.24	87	0.27	0.27	0.27	0.25	68
KL	2	0.24	0.24	0.25	0.23	79	0.26	0.26	0.28	0.24	60
KL	3	0.24	0.24	0.25	0.23	79	0.26	0.26	0.28	0.24	59
KL	4	0.23	0.24	0.26	0.23	79	0.26	0.26	0.28	0.24	59
KL	5	0.24	0.24	0.26	0.23	83	0.27	0.26	0.28	0.24	62
KL	6	0.27	0.24	0.27	0.24	90	0.29	0.26	0.29	0.25	71

Note: R3 represent SE with precision level 0.25; R6 represent PSER with precision level 0.25;

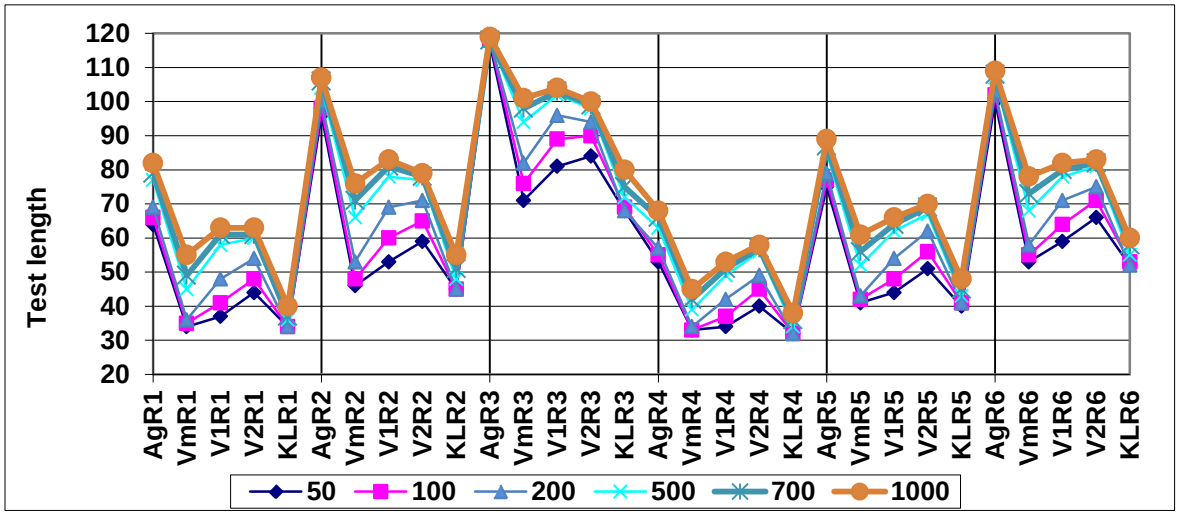


Figure 1: Average test length for the first 50, 100, 200, 500, 700, and 1000 examines for the five procedures, three precision levels, and two stopping rules.

Note for the format for the label of x-axis: XZ--X represents method Ag, Vm, V1, V2, KL; Y represents precision level 0.35, 0.3, and 0.25 and stopping rule SE and PSER; SE: R1-R3; PSER: R4-R6.

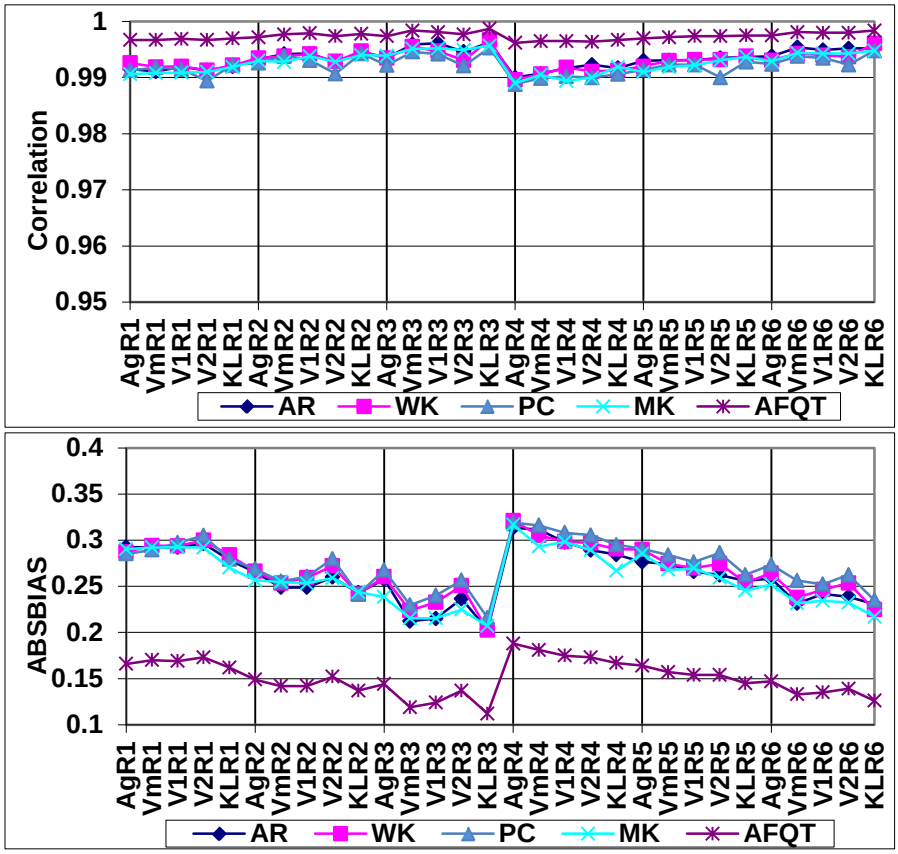
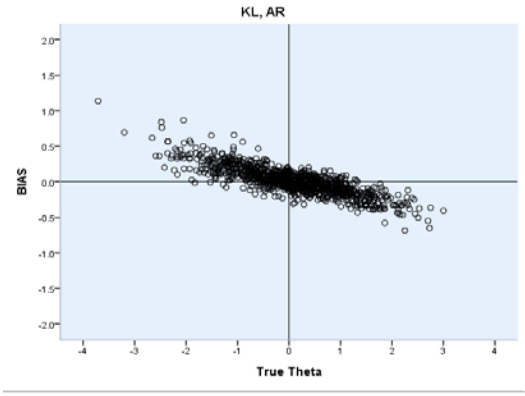
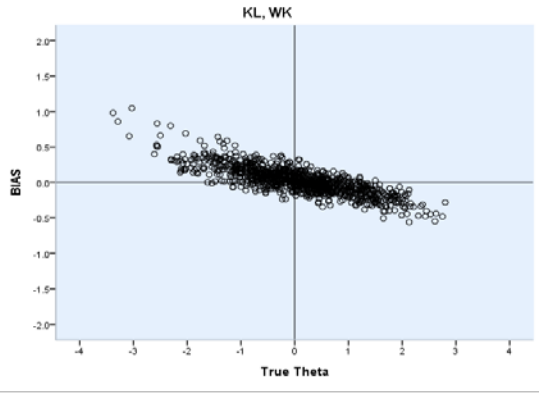
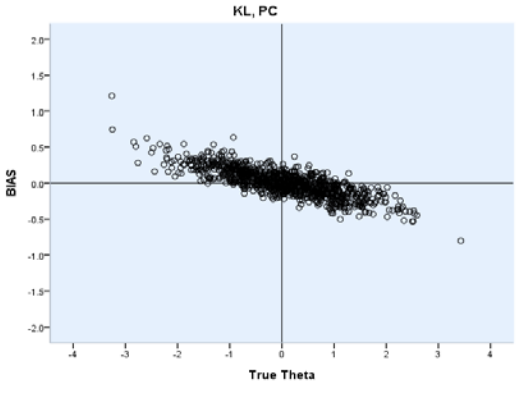


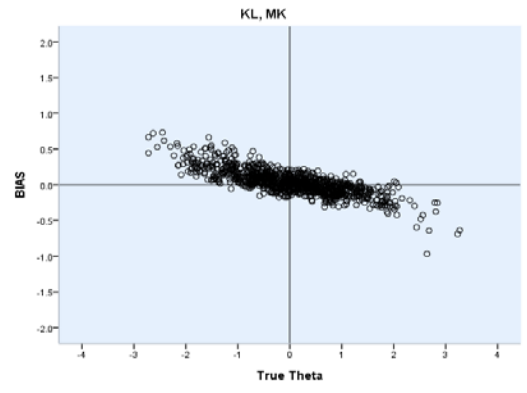
Figure 2: Correlations and ABSBIAS for the four domains and overall scores for the five procedures, three precision levels, and two stopping rules.

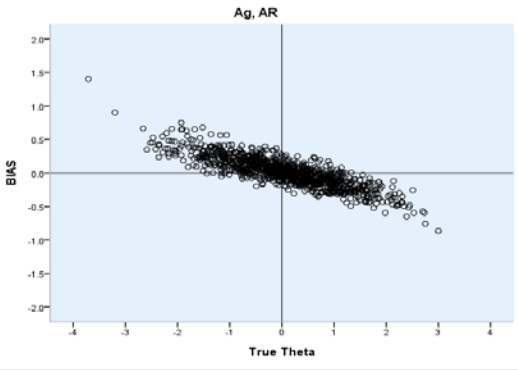
Note for the format for the label of x-axis: XZ--X represents method Ag, Vm, V1, V2, KL; Y represents precision level 0.35, 0.3, and 0.25 and stopping rule SE and PSER; SE: R1-R3; PSER: R4-R6.

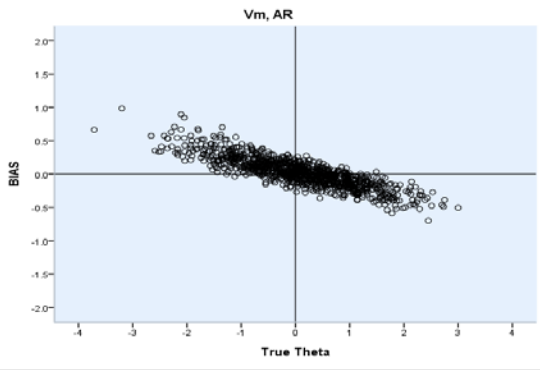

KL, AR
BIAS
True Theta

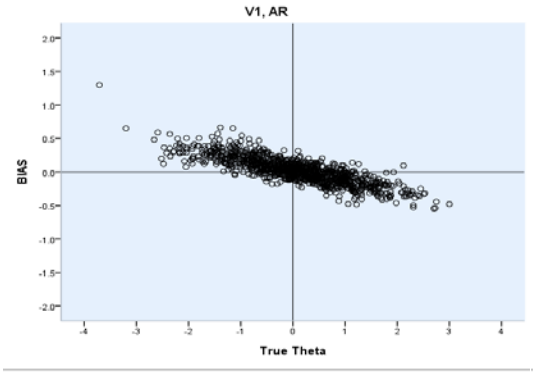

KL, WK
BIAS
True Theta

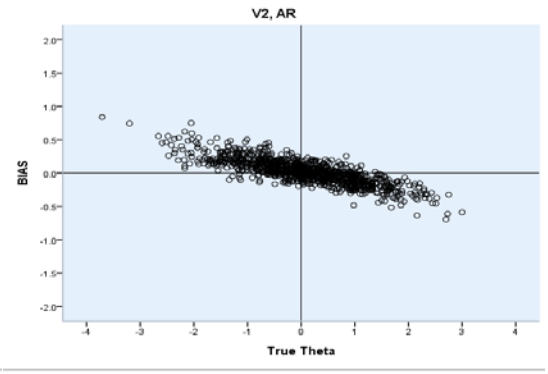

KL, PC
BIAS
True Theta


KL, MK
BIAS
True Theta


Ag, AR
BIAS
True Theta


Vm, AR
BIAS
True Theta


V1, AR
BIAS
True Theta


V2, AR
BIAS
True Theta

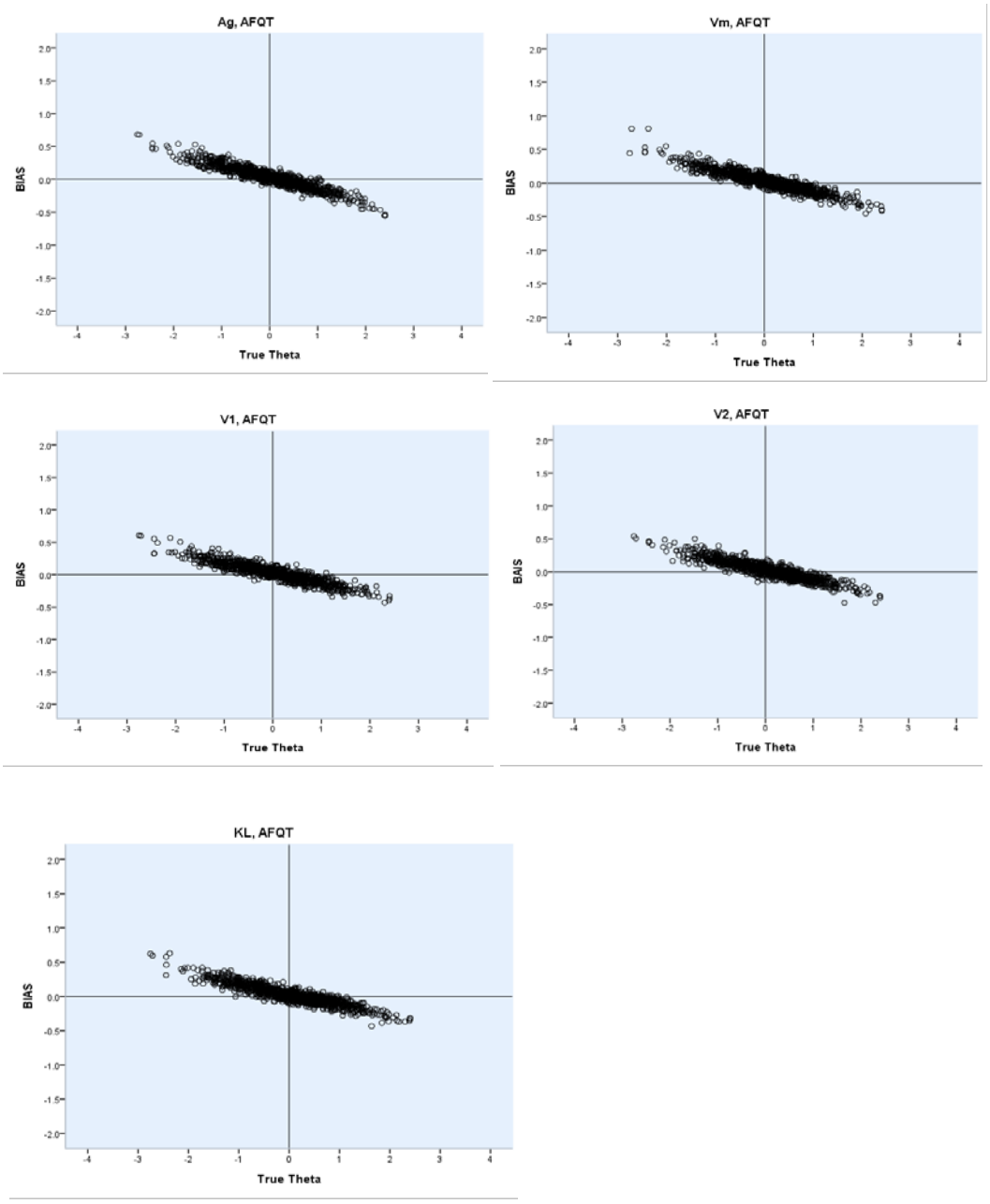


Figure 3: BIAS against the true ability values for the five selection methods for domains and overall for condition PSER with precision level .35 (R4).

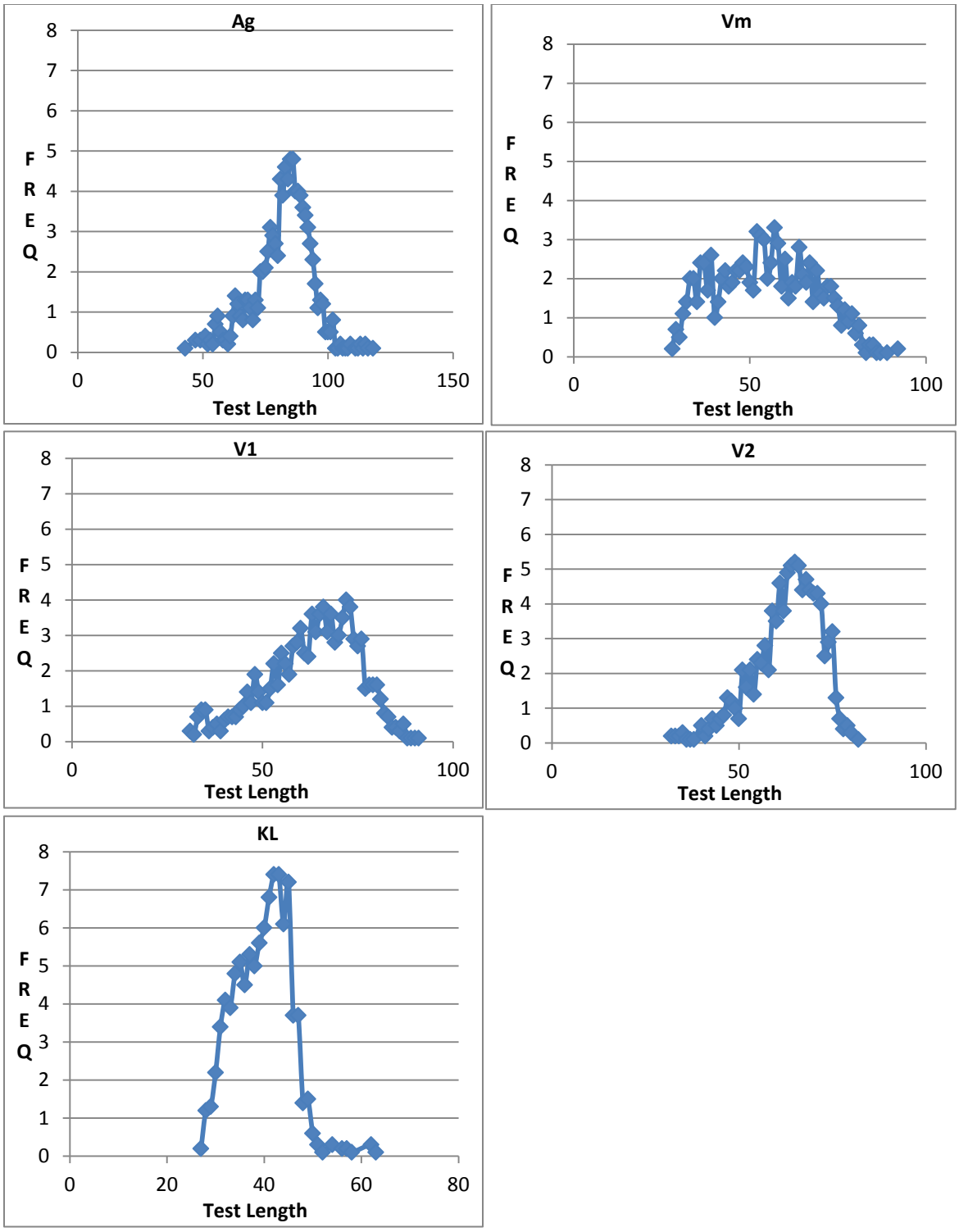


Figure 4: Distributions of the test lengths for the 1000 examinees for the five selection methods for condition R1 with SE stopping rule and .35 as the precision.