Assessing the efficiency of item selection in computerized adaptive testing

Alexander Weissman

Law School Admission Council

Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL, April 2003. Please direct correspondence to Alexander Weissman, Law School Admission Council, 661 Penn Street, Newtown, PA 18940. Email: aweissman@lsac.org.

Abstract

This study investigated the efficiency of item selection in a computerized adaptive test (CAT), where efficiency was defined in terms of the accumulated test information at an examinee's true ability level. A simulation methodology compared the efficiency of two item selection procedures with five ability estimation procedures for CATs of 5-, 10-, 15-, and 25-items in length. The two item selection procedures included maximum Fisher information (FI) and maximum Fisher interval information (FII) item selection. The five ability estimation procedures included maximum likelihood (ML), modal a posteriori (MAP), golden section search (GSS), and two new procedures proposed in this study. These procedures, ML/Alt and MAP/Alt, adjusted ML or MAP estimates according to a specific decision rule based on hypothesis-testing.

For the conventional item selection procedure (FI) and ability estimation procedures (ML and MAP), the best performance was observed for FI with MAP at middle ability levels, with efficiency attaining or exceeding 90% even for the shortest test length. In contrast, larger gaps in efficiency were observed for FI with MAP at extreme ability levels, and for FI with ML across all ability levels. Utilizing FII item selection with ML and MAP narrowed the gaps in efficiency at the lowest ability levels for 5- and 10-item tests. The greatest increase in test efficiency was observed when the alternative ability estimation procedures (ML/Alt, MAP/Alt, and GSS) were used. The gains in efficiency were most pronounced for shorter tests, but were noticeable even for longer tests. Overall, it appears that ability estimation procedure impacts the efficiency of item selection to a larger extent than item selection procedure.

1 Introduction

Efficiency is often cited as an advantage of computerized adaptive tests (CATs) over traditional paper-and-pencil tests. Typically, a CAT version of a test requires half as many items to be administered as its paper-and-pencil counterpart, without compromising measurement precision (Stocking, Smith & Swanson, 2000). Nevertheless, the efficiency of a CAT at the early stages of test administration has been a point of contention in the literature. At the early stages of a CAT administration, provisional ability estimates are typically imprecise, inaccurate, or both. Because item selection is dependent on ability estimation, the arguments contend that item selection based on these early provisional ability estimates is likely to be mismatched with respect to an examinee's true ability. Chen, Ankenmann, and Chang (2000) point out that the inaccuracy of these provisional ability estimates early in CAT administration is "a persistent problem" and that "the more accurate [the provisional ability estimate] is, the more appropriate the selected item will be."

The recognition that provisional ability estimates at the early stages of testing are inaccurate has generated an area of research which seeks to improve the efficiency of a CAT by means of alternative item selection procedures and alternative ability estimation procedures. Recent studies examining the efficacy of alternative item selection procedures suggest that all perform similarly to each other as well as to FI item selection after ten items have been administered (Chen, Ankenmann, & Chang, 2000; Cheng & Liou, 2000). Although it is perhaps unlikely that a CAT of 10 or less items would be administered operationally, the question remains as to whether the efficiency of a CAT might be improved at the early stages of administration by perhaps another item selection

or ability estimation procedure not yet considered, and that such potential gains in efficiency obtained early on might translate into more precise measurements after considerably more items have been administered.

It should be noted that almost all research on improving the efficiency of CAT item selection has concentrated on alternative item selection procedures. However, ability estimation plays an equally important role in CAT item selection, as any item selection procedure must utilize provisional ability estimates. Xiao (1999) demonstrated that an alternative ability estimation procedure utilizing a golden section search (GSS) optimization technique was as accurate as the more common expected a posteriori (EAP) ability estimation procedure in classifying examinees in a computerized adaptive classification test.

A related issue is the precise meaning of the term "efficiency" and how it should be measured. In studies by Chang & Ying (1996), Chen, Ankenmann, & Chang (2000), and Cheng & Liou (2000), it appears that efficiency is defined in terms of the appropriateness of a selected item with respect to an examinee's true ability. By this definition, therefore, efficient item selection is characterized by the selection of items appropriate to an examinee's true ability. Nevertheless, all of these studies use as outcome measures characteristics of the ability estimates (e.g., root-mean-square errors, bias, and standard errors), as opposed to the characteristics of the selected items themselves.

Davey (2002, personal communication) suggests that a less confounded outcome measure is accumulated test information at an examinee's true ability θ. This measure is calculated on the basis of the items selected for administration, and does not incorporate

errors in ability estimation.[1] Through this measure, a precise definition of efficiency may be obtained, one that follows naturally from the statistical concepts of efficiency and relative efficiency.

The objectives of the present study are then: (1) define precisely the efficiency of item selection in a CAT; (2) quantify the efficiency (or inefficiency) of item selection when conventional item selection and ability estimation procedures are utilized; (3) propose a new alternative ability estimation procedure that addresses potential inefficiencies in CAT item selection; (4) quantify the efficiency of item selection under alternative item selection procedures, alternative ability estimation procedures, or both; and (5) examine the extent to which these alternative configurations improve upon the efficiency of item selection over the conventional procedures.

2 Theoretical framework

2.1 Defining the efficiency of item selection in CAT

Consider two tests, A and B, administered to an examinee possessing true ability θ. The precision with which this examinee may be measured by test A is given by the accumulated test information at the examinee's true ability θ, or $I_A^{(T)}(\theta)$. Likewise, $I_B^{(T)}(\theta)$ indicates the precision afforded by test B. The relative efficiency of test A over test B, indicated by $RE(A,B|\theta)$, is the ratio $I_A^{(T)}(\theta)/I_B^{(T)}(\theta)$. Thus, if test A is more efficient than test B, $RE(A,B|\theta)>1$.

This definition of relative efficiency may be extended to the CAT context, yielding an operational definition for the efficiency of a CAT. Suppose that a CAT of *j*

[1] There can be no question that the specific items selected by the CAT are influenced by the ability estimation method; however, this measure is a function defined only in terms of item parameters and a given value of true ability.

items is administered to an examinee possessing true ability θ, and that these items are drawn from an item bank of finite size. Then the quantity $I_{CAT}^{(T)}(\theta)$ characterizes the accumulated test information from these *j* items at the examinee's ability level. Now for any given θ, there exists an optimal set of items, also of size *j*, such that no other combination of *j* items yields a greater measure of accumulated test information. Thus, if $I_0^{(T)}(\theta)$ represents the accumulated test information for this optimal set of items, the relative efficiency of the set of items selected by the CAT administration over the optimal set is $I_{CAT}^{(T)}(\theta)/I_0^{(T)}(\theta)$. Noting, however, that $I_0^{(T)}(\theta)$ places an upper bound on the precision with which an examinee with true ability θ may be measured by a set of *j* items drawn from the item bank, it must be the case that $I_{CAT}^{(T)}(\theta)/I_0^{(T)}(\theta) \leq 1$. It is this ratio that operationally defines the efficiency of a CAT in the present context.

2.2 Item selection procedures

Of the two item selection procedures considered in this study, one is conventional (maximum Fisher information), the other is alternative (maximum Fisher interval information). Maximum Fisher information (FI) item selection is taken here to be the process whereby: (1) an examinee's provisional ability estimate $\hat{\theta}_j$ is obtained after the j^{th} item has been administered; and (2) the j+1th item is selected such that it both possesses maximum Fisher information at the provisional ability estimate and has not already been administered. Item selection by Fisher interval information (FII) is closely related to maximum FI item selection, but instead of evaluating item information at a single point (i.e., the provisional ability estimate), an information index is evaluated instead (Veerkamp & Berger, 1997). This index is obtained by performing a

mathematical integration of the information function associated with an item along a specified interval of the ability continuum.

2.3 Ability estimation procedures

Of the five ability estimation procedures in the study, two are conventional; namely, maximum likelihood (ML) and maximum a posteriori (MAP) estimation. ML estimation finds the ability estimate $\hat{\theta}_{ML,j}$ that maximizes the likelihood function for an examinee's responses to *j* administered items. MAP estimation finds the ability estimate $\hat{\theta}_{MAP,j}$ that occurs at the maximum of the posterior density function after *j* items have been administered, where the posterior density is proportional to the likelihood multiplied by the prior density, taken here to be $N(0,1)$. The alternative ability estimation procedures—Xiao's (1999) golden section search (GSS) strategy and the proposed alternative procedure—utilize hypothesis-testing. Xiao (1999) obtains provisional ability estimates $\hat{\theta}$ by a golden-section search (GSS) strategy; the next item is selected based on this most current provisional ability estimate. Using GSS, a starting estimate $\hat{\theta}_1$ is identified as the midpoint of a search interval along the ability continuum; a hypothesis test is conducted by comparing optimally-weighted observed and expected scores given $\hat{\theta}_1$ (see Birnbaum, 1968 for a discussion on optimally-weighted scores). If the hypothesis test results in rejection, then a new search interval is identified, as well as a new estimate $\hat{\theta}_2$. The search strategy continues until the null hypothesis is not rejected. The last estimate $\hat{\theta}$ obtained is then taken as the provisional ability estimate.

The proposed alternative ability estimation procedure operates concurrently with a conventional ability estimation procedure such as ML and MAP, yielding two more

alternative ability estimation procedures, denoted as ML/Alt and MAP/Alt. Like Xiao (1999), the alternative procedure conducts a hypothesis test after the j^{th} item in the test has been administered. However, the null hypothesis in the procedure is that all j items administered to an examinee are maximally informative at that examinee's true ability θ; failure to reject the null suggests that the ability estimate obtained by ML or MAP should be used for the subsequent selection of the $j+1^{th}$ item, while rejection of the null suggests a modified ability estimation procedure. This modified ability estimate is found using the expected proportion correct under the null hypothesis, its confidence limits, and the average item characteristic curve for the j administered items.

Here, the null hypothesis is constructed under strict model assumptions. These assumptions follow from the IRT model in the case where all items administered to an examinee are maximally discriminating (i.e., possess maximum information) at that examinee's true ability. Such a scenario characterizes ideal item selection in a CAT; namely, that items administered to an examinee should possess maximum measurement precision at that examinee's true ability. Thus, the hypothesis-testing procedure used here is essentially a test of whether the CAT is operating as intended. In brief, if this null hypothesis is not rejected, then the decision is to use the most recent provisional ability estimate obtained by a conventional ability estimation procedure (e.g., ML or MAP) to select the next item. If evidence warrants its rejection, however, an alternative selection method is suggested. Thus, the alternative procedure functions concurrently with a conventional ability estimation procedure such as ML or MAP, and in this sense acts as an adjustment to the conventional ability estimate when model assumptions do not conform to the observed data.

The overall rationale for this hypothesis-testing procedure is that when a CAT is targeting items exactly at an examinee's true ability, the expected proportion of items correctly answered is approximately equal to 0.5 in the case of items modeled under the 3P IRT model, and is exactly equal to 0.5 in the case of 1P and 2P items. The presence of a pseudo-guessing parameter *c* in the 3P IRT model increases the expected proportion correct from 0.5 to a higher number, with larger values of *c* corresponding to higher expected proportions correct. After an examinee has responded to an administered item, the hypothesis-testing procedure compares the observed proportion of correct responses with what would be expected if the CAT was selecting items perfectly targeted to an examinee's ability. If the observed proportions correct are less than expected, the interpretation is that the current ability estimate is too high. Alternatively, if the observed proportions correct are greater than expected, the interpretation is that the current ability estimate is too low. Thus, a new adjusted ability estimate may be introduced in order to compensate for the discrepancy. It should be noted that the expected proportion correct under this ideal situation may be calculated without knowledge of examinee ability, as will be discussed shortly.

The assumptions underlying the null hypothesis for this procedure are rooted in how IRT characterizes item information. Under the 1, 2, and 3-parameter models, the probability of correct response $P(U_i = 1|\theta)$ for an item *i* is modeled as a monotonically increasing function of θ. However, for each curve suggested by this function, there exists exactly one point where its first derivative is at a maximum. It is also at this point where the item possesses maximum information. Thus, if $\theta_{max,i}$ represents the value on the

ability scale corresponding to this point, then the item possess maximum measurement precision for an examinee whose own true ability θ is equal to $\theta_{\max,i}$.

Now suppose that a set of *N* items are administered to an examinee with true ability θ, and impose the restriction that for each item *i*, $\theta_{\max,i} = \theta$. That is, all *N* items possess maximum information at the examinee's true ability θ. (Note, however, that there is no restriction that all items be equally informative, so it is permissible that $I_i(\theta) \neq I_j(\theta)$ for $i \neq j$.) Thus, in this situation where all items are ideally suited for this examinee in terms of measurement precision,

$$\theta_{\max,1} = \theta_{\max,2} = \cdots = \theta_{\max,N} = \theta \qquad \text{(Eq. 1)}$$

The next relationship links Equation 1 with the statement of the null hypothesis employed by this procedure. Since there is a probability $P(U_i = 1|\theta_{\max,i})$ associated with each item *i*, an expected proportion correct may be constructed, under the constraints imposed by Equation 1. This expected proportion correct, or p, is then defined as

$$p = \frac{\sum_{i=1}^{N} P(U_i = 1|\theta_{\max,i})}{N} \qquad \text{(Eq. 2)}$$

The observed proportion correct, or $\hat{p}$, is defined as

$$\hat{p} = \frac{\sum_{i=1}^{N} X_i}{N}, \quad X_i = \{0,1\} \qquad \text{(Eq. 3)}$$

where $X_i = 0$ indicates an incorrect response and $X_i = 1$ indicates a correct response.

The null hypothesis is then that $\hat{p}$ is sampled from a distribution with mean *p*. Thus, a decision not to reject the null hypothesis implies that the observed proportion

correct does not differ from the expected proportion correct *p*. Because an examinee's ability is assumed fixed at some true value θ, this decision further suggests that the relationship in Equation 1 be retained[2]. In this case, the model would fit the data.

However, if the null hypothesis is rejected, then an alternative hypothesis is required. Rejection of the null implies that the observed proportion correct is inconsistent with what would be expected under Equation 1; that is, a discrepancy must therefore exist between the $\theta_{\max,i}$ for the $i = \{1, 2, \ldots, N\}$ items administered and that examinee's true ability θ. Thus, the model does not fit the data.

In order to conduct the necessary hypothesis tests, a test statistic and its distribution is required. To begin, consider an examinee's dichotomous response X_i to item *i*. Then according to the IRT model, $X_i \sim \text{BIN}(1, p_i)$, such that X_i is a Bernoulli random variable with parameter p_i, and the parameter $p_i = P(X_i = 1|\theta)$ for constant θ. Now assume that a sample of size *n* is taken, where the X_i are independent but not identically distributed. (Thus, local independence of item responses is assumed here.) The proportion correct for X_i (or, the mean of the X_i) may then be defined as

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad \text{(Eq. 4)}$$

[2] If items are perfectly targeted at examinee ability, then Equation 2 follows by deduction. However, the inductive step is somewhat more involved. Satisfying Equation 2 is a necessary but not sufficient condition for concluding Equation 1. Caution must be exercised in interpreting model-data fit under retention of the null hypothesis. Nevertheless, if Equation 2 is not satisfied (i.e., when the null is rejected), it cannot be the case that Equation 1 is true.

Now the expectation $E[\hat{p}]$, denoted by *p*, is

$$p = E[\hat{p}] = E\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{1}{n} E\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{\sum_{i=1}^{n} p_i}{n} \qquad \text{(Eq. 5)}$$

since for $X_i \sim \text{BIN}(1, p_i)$, $E[X_i] = p_i$. The variance of $\hat{p}$, denoted by $Var[\hat{p}]$, is

$$Var[\hat{p}] = Var\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{1}{n^2} Var\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var[X_i] = \frac{\sum_{i=1}^{n} p_i q_i}{n^2} \qquad \text{(Eq. 6)}$$

since the variance of the sum of independent random variables is equal to the sum of their variances, and $Var[X_i] = p_i q_i$, where $q_i = 1 - p_i$.

The test statistic is constructed as

$$z^* = \frac{\hat{p} - p}{\sqrt{Var[\hat{p}]}} \qquad \text{(Eq 7)}$$

where, under the null hypothesis, z^* is asymptotically normally distributed with mean 0 and variance 1, that is, $z^* \xrightarrow{d} N(0,1)$.

For utilizing this hypothesis-testing procedure in the CAT environment, the quantities *p* and $Var[\hat{p}]$ from Equations 5 and 6 are calculated based on the items administered to the examinee, with the assumption under the null hypothesis that all items possess maximum information at the examinee's true ability, as given by Equation 1. Thus, under the 3P model, the p_i for an item *i* used in these equations is given by

$$p_i = P\left(X_i = 1 \middle| \theta_{\max,i}\right) = c_i + \frac{\left(1 - c_i\right)}{\left[1 + \exp\left(-Da_i\left(\theta_{\max,i} - b_i\right)\right)\right]} \qquad \text{(Eq 8)}$$

where $\theta_{\max,i}$ is directly attainable from the item parameters for item *i*, and is given by (Hambleton and Swaminathan, 1985)

$$\theta_{\max,i} = b_i + \frac{1}{Da_i} \ln\left(\tfrac{1}{2} + \tfrac{1}{2}\sqrt{1 + 8c_i}\right) \qquad \text{(Eq 9)}$$

and the a_i, b_i, and c_i are the discrimination, difficulty, and pseudo-guessing parameters, respectively, for item *i*; *D* is a scaling constant. Substituting the expression for $\theta_{\max,i}$ from Equation 9 into Equation 8 results in the following simplification for p_i

$$p_i = P\left(X_i = 1 \middle| \theta_{\max,i}\right) = c_i + \left(1 - c_i\right)\left[1 + \frac{2}{1 + \sqrt{1 + 8c_i}}\right]^{-1} \qquad \text{(Eq 10)}$$

Using this expression for p_i, the necessary quantities $E\left[\hat{p}\right]$ and $Var\left[\hat{p}\right]$ may be calculated by means of Equations 5 and 6.

Equation 7 is used to test the null hypothesis that all items administered to an examinee are maximally informative at that examinee's true ability θ. If the absolute value of the test statistic z^* exceeds a critical value z_c, then the null hypothesis is rejected. Otherwise, the null hypothesis is retained. The provisional ability estimate used for selecting the next item depends on this decision rule.

Null hypothesis not rejected. In instances where the null hypothesis is not rejected (i.e., $\left|z^*\right| \le z_c$), there is not sufficient evidence to suggest that items are not maximally informative at an examinee's ability θ. The recommendation therefore is that

the most recently-obtained provisional ability estimate (from ML or MAP, for example) be used to select the next item.

Null hypothesis rejected. Sufficient evidence warrants the rejection of the null hypothesis in this case (i.e., $|z^*| > z_c$). Selection of the next item based on the most recently-obtained provisional ability estimate is not recommended, and so an alternative ability estimate is suggested. A new provisional ability estimate $\hat{\theta}^*$, different from that estimated either by ML or MAP, is thus identified. This estimate is found using the expected proportion correct *p*, its confidence limits under the null hypothesis, and the average item characteristic curve for the administered items. Item selection then proceeds based on this new provisional estimate $\hat{\theta}^*$.

In this case where the null hypothesis is rejected, it is concluded that the sample proportion correct $\hat{p}$ is not from a distribution with mean *p*. Since the hypothesis test is constructed under the null hypothesis, inference does not extend to the distribution from which $\hat{p}$ is sampled. That is, the hypothesis test alone cannot characterize the alternative mean of $E[\hat{p}]$. However, a conservative estimate of the location of this alternative distribution is possible.

Let p_0 denote the expected proportion correct under the null hypothesis, and p_α the expected proportion correct under the alternative. At the very least, the alternative distribution becomes distinguishable from the null distribution at the decision threshold; that is, at either one of the confidence limits set for p_0. Thus, a decision to reject the null hypothesis when $\hat{p} < p_0$ is equivalent to stating that $\hat{p}$ lies outside the confidence interval for p_0, and specifically, beyond its lower confidence limit of $p_0 - z_c\sqrt{Var[p_0]}$.

Likewise, rejection of the null when $\hat{p} > p_0$ demands that $\hat{p}$ must lie beyond the upper confidence limit $p_0 + z_c\sqrt{Var[p_0]}$.

Thus, an approximation to p_α may be denoted by $\hat{p}^*$, such that

$$\begin{aligned} \hat{p}^* &= p_0 + z_c\sqrt{Var[p_0]}, \quad \hat{p} > p_0 \\ \hat{p}^* &= p_0 - z_c\sqrt{Var[p_0]}, \quad \hat{p} < p_0 \end{aligned} \qquad \text{(Eq. 11)}$$

where each of the quantities p_0, $Var[p_0]$, and z_c are as defined under the hypothesis-testing procedure.

By itself, the estimate $\hat{p}^*$ is not particularly useful for identifying a new provisional ability estimate, since it is a proportion, not a value on the ability scale. However, the average item characteristic curve (ICC) provides a means for relating proportions to ability values. Through the average ICC, the $\hat{p}^*$ obtained from the hypothesis-testing procedure may be converted to a new provisional ability estimate $\hat{\theta}^*$. The use of the average ICC in such a manner is justified under the IRT model, since the probabilities associated with a correct response for a given item are dependent only on examinee ability θ.

The average of the ICCs from all administered items, or the average ICC, is equivalent to the test characteristic curve (TCC) divided by the number of items administered. Because an analytical solution is not available to transform $\hat{p}^*$ to $\hat{\theta}^*$ through the average ICC, a numerical search procedure is required. The procedure uses the method of halving, where a discrete interval $[a,b]$ is halved at each iteration, producing a midpoint $c = (a+b)/2$. The average ICC function $\overline{P}(X = 1|\theta)$, defined as

$$\overline{P}\left(X=1|\theta\right)=\frac{\sum_{i=1}^{N}P\left(X_i=1|\theta\right)}{N} \qquad \text{(Eq. 12)}$$

for items 1, 2, …, N is then evaluated at $\theta=\{a,c,b\}$. If $\hat{p}^*$ is within the interval $[a,c]$, that is, when $\hat{p}^* \le \overline{P}\left(X=1|\theta=c\right)$, then the interval boundary points are updated to be $[a,c]$ for the next iteration. Otherwise, $\hat{p}^*$ is within the interval $[c,b]$ and the interval boundary points are updated as $[c,b]$. This method of halving continues until the maximum number of iterations has been met. For this study, the lower bound on ability was set at $\theta=-4$, the upper bound at $\theta=+4$, and the maximum number of iterations for the method of halving was set to 15.

Recall that the alternative ability estimation procedure employs a critical *z*-value for hypothesis-testing. In many applications, the critical *z*-value is set beforehand to correspond to a nominal α-level, such as $z_c = 1.96$ for $\alpha = 0.05$, in order to control the Type I error rate. However, in the context of the alternative ability estimation procedure, a decision to set α to a small value (such as 5%) translates into infrequent invocation of the procedure, and hence the hypothesis test may be too conservative. What is required is a method for determining an optimal value of z_c that will allow the alternative procedure to function more frequently while maximizing correct decisions and minimizing incorrect decisions.

Two optimal z_c values were determined empirically, one for Alt/ML estimation and the other for Alt/MAP estimation. The values were found by conducting simulations under these procedures and examining two measures: (1) the accuracy of the $\hat{\theta}^*$ alternative ability estimates with respect to examinee true ability; and (2) the relative

efficiency of tests administered using the alternative procedures (i.e., Alt/ML or Alt/MAP) as compared to tests administered using the corresponding conventional procedures (i.e., ML or MAP). Maximum FI item selection was used for all simulations, and the item pool used for these simulations was the same as that used for the full study. It was found that $z_c = 0.9$ was optimal for Alt/ML, and $z_c = 1.3$ was optimal for Alt/MAP. While these critical *z*-values may appear small, it should be noted that Xiao (1999), for her hypothesis-testing procedure, suggested a critical value of $z = 0.7$; this value was also determined empirically.

3 Experimental design

This study employed a CAT simulation methodology; the simulations used an item bank of 367 pre-calibrated and dichotomously-scored 3-parameter IRT items from a recently-administered large-scale CAT assessment of mathematics ability. The four factors in the fully-crossed experimental design were: (1) item selection procedure (maximum FI or maximum FII item selection); (2) ability estimation procedure (ML, MAP, GSS, ML/Alt, or MAP/Alt); (3) true ability level at discrete points along the ability continuum (at $\theta = \{-2, -1, 0, +1, +2\}$); and (4) test length (5, 10, 15, or 25 items). For each of the experimental conditions, 1000 replications were generated. The layout of the experimental design is given in Table 1.

Efficiency, as defined earlier, is the primary dependent measure. Since analyses indicate that this measure is highly skewed to the left, the median efficiency is reported as a measure of central tendency, and the interquartile range is reported as a measure of variability.

[Insert Table 1 about here]

4 Results

One objective of this research was to quantify the efficiency (or inefficiency) of item selection when conventional item selection and ability estimation procedures were utilized. The efficiency measure $I_{CAT}^{(T)}(\theta)/I_0^{(T)}(\theta)$ helped to address this question, as it indicated how efficient a given procedure was with respect to the maximum efficiency attainable. Table 2 provides efficiency measures for the conventional item selection procedure (FI) and ability estimation procedures (ML and MAP). Under maximum FI item selection, MAP was more efficient than ML at the middle ability levels $\theta = \{-1, 0, 1\}$, and less efficient than ML at the extreme ability levels $\theta = \{-2, 2\}$ for all tests lengths (5, 10, 15, and 25 items), although these differences became smaller as test length increased.

[Insert Table 2 about here]

The quantification of efficiency indicates how well, in terms of optimal performance, the procedures are operating. As shown in Table 2, while ML was indeed more efficient than MAP at the extreme ability levels, median efficiencies at these ability levels did not exceed 62% for 5 items, and did not exceed 82% for 10 items. In contrast, at the middle ability levels where MAP was more efficient, MAP efficiencies exceeded 88% for 5 items, and 91% for 10 items. Thus, one finding here is that little room for improvement exists for maximum FI item selection with MAP ability estimation at middle ability levels, as it attained nearly 90% or greater efficiency even for the shortest test length. Where room for improvement does exist is for ML ability estimation, across all levels of ability, and for MAP at the extremes. For both of these cases, the largest gaps in performance occurred for the shorter test lengths.

It was hypothesized that alternative item selection procedures, alternative ability estimation procedures, or a combination of both might prove useful for narrowing the gaps in efficiency observed under conventional procedures. The extent to which each of these alternative configurations might improve upon the efficiency of item selection over the conventional procedures is now examined.

4.1 Alternative item selection with conventional ability estimation

One possibility for narrowing the gaps in efficiency is to utilize an alternative item selection procedure, but maintain conventional ability estimation. Maximum FII item selection, an alternative procedure, was examined in conjunction with ML and MAP ability estimation. As shown in Table 3, under maximum FII item selection, the performance of MAP and ML is enhanced at the extreme ability levels for short tests, but no change is observed for longer tests. Interestingly, some of these results are consistent with prior research; e.g., Chen, Ankenmann, & Chang (2000). Although Chen et al.'s (2000) dependent measures were different from those utilized here (bias, standard error, and RMSE of ability estimates versus efficiency measures) and ability estimation procedure was different (EAP versus ML and MAP), they also found that maximum FII item selection performed better than maximum FI item selection at the lower extreme of ability ($\theta = -2$) for tests 10 items in length or shorter.

[Insert Table 3 about here]

In the present study it was found that in addition to increased efficiency at the lower extreme of ability, FII item selection benefited MAP estimation (but not ML) at the higher extreme of ability ($\theta = 2$), for the 5- and 10-item tests. Maximum FII item selection raised median efficiency measures in the case of MAP by about 10% for 5-item

tests, and 6% for 10-item tests. The greatest increase in median efficiency under maximum FII selection was observed for ML at the lowest ability level, with an increase of 30% over maximum FI selection at 5 items.

4.2 Conventional item selection with alternative ability estimation

Another possibility for narrowing the gaps in efficiency is to maintain conventional item selection, but utilize an alternative ability estimation procedure. The efficiency measures from the alternative ability estimation procedures ML/Alt, MAP/Alt, and GSS under maximum FI item selection are provided in Table 2. In general, the alternative procedures ML/Alt and MAP/Alt helped fill the gaps in the efficiency of the conventional ML and MAP procedures under maximum FI item selection, without negatively impacting them in cases where performance was already high. The alternative ability estimation procedures yielded higher median efficiency measures while simultaneously maintaining or decreasing variability in those measures. The improvement in efficiency was greater than that observed for ML and MAP under maximum FII selection, and occurred across more ability levels. For instance, ML estimation only benefited from maximum FII selection at $\theta = -2$, whereas efficiency measures for ML/Alt were higher for all ability levels. Further, while maximum FII selection did augment the median efficiency of 5- and 10-item tests at $\theta = -2$ for ML estimation under maximum FI selection by 30% and 8%, respectively, ML/Alt saw a corresponding increase of 47% and 20%, respectively, under maximum FI selection.

Both ML/Alt and MAP/Alt were new methods proposed in this study. However, the GSS ability estimation procedure had been previously investigated by Xiao (1999). As shown in Table 2, median efficiency measures from GSS are always higher than those

from ML, and the differences are most pronounced for shorter test lengths. Interestingly, results from the GSS procedure closely parallel those from ML/Alt. This correspondence may result from the fact that GSS, like ML/Alt, utilizes hypothesis-testing and an interval search strategy.

4.3 Alternative item selection with alternative ability estimation

Yet another possibility for narrowing the gaps in efficiency is to utilize both alternative item selection and ability estimation procedures. The efficiency measures from the alternative ability estimation procedures ML/Alt, MAP/Alt, and GSS under maximum FII item selection are provided in Table 3. Under maximum FII, the alternative ability estimation procedures again narrow the gaps in efficiency observed for ML and MAP. However, there is no clear performance advantage for using the alternative ability estimation procedures under maximum FII selection as opposed to maximum FI selection. The results were mixed for 5- and 10-item tests, and were essentially unchanged for longer test lengths. Two median efficiency measures were lower under maximum FII item selection for 5-item tests; they occurred for ML/Alt and GSS at $\theta = 2$. One measure was higher, also for ML/Alt but at $\theta = 0$. No clear pattern for the change in variability measures was observed. In the nine cases where differences in variability were detected, three were increases.

5 Discussion

Overall, it appears that ability estimation procedure impacts the efficiency of item selection to a larger extent than item selection procedure. The effect of alternative ability estimation procedures (ML/Alt, MAP/Alt, and GSS) on test efficiency was greater than the effect of the alternative item selection procedure (FII). Thus, incorporating ability

estimation error into item selection procedures (as is the case with alternative item selection procedures such as FII) may be less effective at increasing test efficiency than utilizing alternative ability estimation procedures.

Item selection and ability estimation are two necessary ingredients for a CAT. However, improvements in one area may be offset by weaknesses in the other. The present study attempts to isolate the effects of item selection on efficiency by utilizing an outcome measure that is not confounded by ability estimation. In addition, the proposed upper bound on efficiency is independent of the particular ability estimation employed and serves as the theoretical limit for measurement precision.

While it has been posited that maximum FI item selection with conventional ability estimation procedures is inefficient at the early stages of testing, this study addressed the question, to what *extent* is maximum FI item selection with these ability estimation procedures inefficient? It further addressed the question, what is the utility in employing alternative item selection or ability estimation procedures? The answers to these questions are likely of interest to the measurement practitioner who must assemble CATs for large-scale administration. Alternative item selection and ability estimation procedures that are relatively easy to implement in an operational setting were suggested and evaluated.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick, *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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

Chen, S.-Y., Ankenmann, R.D., & Chang, H.-H. (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement*, *24*, 241-255.

Cheng, P.E. & Liou, M. (2000). Estimation of trait level in computerized adaptive testing. *Applied Psychological Measurement*, *24*, 257-265.

Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory: principles and applications*. Norwell, MA: Kluwer Academic Publishers.

Stocking, M.L., Smith, R., & Swanson, L. (2000, April). *An investigation of approaches to computerizing the GRE subject tests* (Research Report 00-4). Princeton, NJ: Educational Testing Service.

Veerkamp, W.J.J. & Berger, M.P.F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics*, *22*, 203-226.

Xiao, B. (1999). Strategies for computerized adaptive grading testing. *Applied psychological measurement*, *23*, 136-146.

Table 1. Layout of experimental design.

Item selection	Ability estimation	Test length (in items)	True ability θ				
			−2	−1	0	+1	+2
Maximum FI	ML	5, 10, 15, 25					
	ML/Alt	5, 10, 15, 25					
	MAP	5, 10, 15, 25					
	MAP/Alt	5, 10, 15, 25					
	GSS	5, 10, 15, 25					
Maximum FII	ML	5, 10, 15, 25					
	ML/Alt	5, 10, 15, 25					
	MAP	5, 10, 15, 25					
	MAP/Alt	5, 10, 15, 25					
	GSS	5, 10, 15, 25					
Dependent measures provided:			• Efficiency at 50^{th} percentile (median), IQR				

Table 2. Medians and interquartile ranges of the efficiency measure under maximum FI item selection.

Ability estimation	Test length	Median efficiency					Efficiency interquartile range (IQR)				
		θ=-2	θ=-1	θ=0	θ=1	θ=2	θ=-2	θ=-1	θ=0	θ=1	θ=2
ML	5	53.0	73.2	54.2	44.8	61.6	44.3	12.1	29.3	45.4	17.1
	10	81.8	83.9	71.9	70.1	80.7	18.8	18.9	29.0	35.1	13.3
	15	93.0	87.1	80.1	80.3	90.9	9.1	15.9	21.8	22.0	7.1
	25	96.7	92.7	89.5	89.3	95.8	5.5	9.6	12.2	11.6	3.2
ML/Alt	5	100.0	94.3	63.3	83.0	93.9	18.5	7.8	26.7	40.0	16.4
	10	99.5	91.8	81.1	86.2	100.0	11.7	16.5	27.0	30.6	13.8
	15	99.9	92.8	87.4	89.2	99.5	4.0	14.3	21.4	20.6	5.0
	25	99.0	96.1	93.7	94.5	99.6	2.2	8.4	12.6	11.2	2.2
MAP	5	31.0	91.4	88.5	95.9	23.6	49.2	19.6	15.6	23.6	0.0
	10	73.2	94.2	91.7	92.5	64.3	29.1	20.2	14.9	13.6	13.6
	15	87.1	92.9	93.2	94.6	85.4	18.4	16.0	12.4	11.0	9.1
	25	90.8	96.1	97.1	96.7	93.9	8.1	9.1	5.9	7.4	3.1
MAP/Alt	5	79.7	90.2	88.5	90.6	54.6	22.5	14.9	21.3	22.2	0.0
	10	81.8	92.3	91.7	90.3	74.5	21.8	19.2	14.5	17.1	19.6
	15	91.4	92.8	93.0	92.8	88.3	19.1	14.7	12.2	13.9	10.2
	25	94.3	96.2	96.5	95.5	94.8	8.7	7.8	6.7	8.3	3.2
GSS	5	96.1	86.5	73.6	81.1	81.1	17.0	20.9	19.8	31.7	2.9
	10	90.9	88.5	78.1	83.6	87.6	15.7	11.8	25.9	31.1	12.3
	15	96.2	89.3	84.4	87.8	95.7	7.2	12.2	20.2	18.4	6.3
	25	97.4	94.8	91.2	94.1	98.3	5.6	9.7	12.1	10.1	2.9

Table 3. Medians and interquartile ranges of the efficiency measure under maximum FII item selection.

Ability estimation	Test length	Median efficiency					Efficiency interquartile range (IQR)				
		θ=-2	θ=-1	θ=0	θ=1	θ=2	θ=-2	θ=-1	θ=0	θ=1	θ=2
ML	5	82.8	74.5	54.2	45.7	65.4	49.9	10.8	29.3	46.3	3.4
	10	89.7	85.7	72.1	72.4	83.3	17.9	14.9	28.0	26.7	13.9
	15	96.2	88.2	80.9	80.4	92.3	8.8	14.9	19.5	19.8	5.8
	25	98.0	93.9	89.2	90.4	96.5	5.5	9.2	11.0	10.4	2.9
ML/Alt	5	100.0	89.9	71.6	82.7	84.8	20.2	13.2	17.1	21.5	15.2
	10	99.3	92.3	81.6	87.3	96.9	12.3	16.0	24.0	24.3	16.3
	15	99.9	92.3	87.2	90.6	98.7	5.0	14.0	20.3	14.9	6.9
	25	99.0	96.2	93.1	95.1	98.6	5.2	8.8	10.7	9.6	2.6
MAP	5	42.8	93.0	90.8	92.4	32.1	49.2	19.9	15.6	17.9	0.0
	10	79.1	91.1	92.0	91.2	70.2	29.3	19.6	12.8	11.2	9.9
	15	89.8	92.1	93.5	94.7	85.9	18.9	15.3	11.1	9.9	7.0
	25	91.2	95.8	96.7	97.0	93.8	9.1	8.5	5.9	6.5	3.2
MAP/Alt	5	79.7	89.5	90.8	92.4	57.7	22.5	12.9	21.3	22.6	0.0
	10	81.9	90.6	91.7	89.7	71.3	27.7	18.9	14.8	11.5	18.5
	15	91.4	92.2	93.2	93.1	87.7	22.1	14.4	11.7	10.8	9.7
	25	94.3	95.9	96.3	96.3	94.3	12.1	8.1	6.4	7.3	4.3
GSS	5	96.1	88.0	73.2	82.7	71.6	17.0	21.2	19.8	21.5	26.8
	10	94.3	86.3	78.1	87.7	89.9	15.7	14.5	26.9	26.1	15.2
	15	96.2	89.8	85.4	88.9	96.3	5.3	11.9	22.3	16.5	6.4
	25	97.4	94.5	91.4	94.6	98.6	5.5	8.8	11.8	9.7	2.5