ASSESSING THE EFFICIENCY OF ITEM SELECTION IN COMPUTERIZED ADAPTIVE TESTING

by

Alexander Weissman

B.A., State University of New York at Potsdam, 1993

M.S., Purdue University, 1995

M.A., Carnegie Mellon University, 1999

Submitted to the Graduate Faculty

of the School of Education in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2002

Faculty Endorsement and

Final Review Committee

Committee Member	Affiliation
Clement A. Stone, Ph.D., Chair	School of Education
Suzanne Lane, Ph.D.	School of Education
Neil H. Timm, Ph.D.	School of Education
Tim Davey, Ph.D.	Educational Testing Service

Date

ASSESSING THE EFFICIENCY OF ITEM SELECTION IN COMPUTERIZED ADAPTIVE TESTING

Alexander Weissman, Ph.D.

University of Pittsburgh, 2002

Adviser: Dr. Clement A. Stone _____

This study investigated the efficiency of item selection in a computerized adaptive test. Efficiency was defined in terms of the accumulated test information at an examinee's true ability level, with a measure of 100% indicating maximally efficient item selection. The study employed a simulation methodology to compare the efficiency of two item selection procedures with five ability estimation procedures for fully-adaptive tests 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 the efficiency of item selection 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.

FOREWORD

I owe a debt of gratitude to a number of people who made the successful completion of this work possible. First, I wish to thank my advisor, Clement Stone, for his support and guidance in seeing this project through, as well as to Suzanne Lane, Neil Timm, and Tim Davey. Their efforts enriched the quality of this research, and I am most grateful for their time and attention.

Special thanks are due to my friends and family, whose unconditional support must certainly be acknowledged. Thanks to Leonardo Hsu, who in addition to being a friend, also provided technical assistance in this research. To my family, I must extend my sincerest appreciation for encouraging me during the challenging times and celebrating the successes. Thanks to my parents, my grandmother, my sisters Becky and Elizabeth, and my brother Greg.

I dedicate this work to my wife, Kathia, whose positive outlook, enduring faith, and genuine care nurtured me throughout this research. Without question, she is my “strength in need, counselor in perplexity, and companion in joy.”

TABLE OF CONTENTS

Page

CHAPTER

LIST OF TABLES

LIST OF FIGURES

Page

Figure

Page

Figure

CHAPTER 1

Introduction

Efficiency of item selection in the context of computerized adaptive testing

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). The CAT administers items targeted to examinee ability, where higher-ability examinees generally receive more difficult items and lower-ability examinees generally receive less difficult items. Under the formulation of item response theory (IRT), it is suggested that much is to be gained in terms of test efficiency by administering items to examinees that are well-targeted to their ability.

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 (i.e., estimates possess large standard errors of measurement), inaccurate (i.e., estimates are biased), 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” (p. 241).

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. While most commonly, maximum information (or Fisher information) item selection is used to select items in a CAT, it has been argued that maximum Fisher information (FI) item selection is inefficient at the early stages of a CAT because it selects items whose information is at the maximum of an inaccurate or imprecise provisional ability estimate as opposed to an examinee’s true ability. Thus, a number of other methods have been proposed which seek either to incorporate the error of ability estimation into item selection (i.e., methods addressing imprecision), or to use a likelihood-ratio based method to identify more suitable items across a range of plausible ability levels (i.e., methods addressing bias). Methods developed under the former approach include the general weighted information criterion (Veerkamp & Berger, 1997) which leads to Fisher interval information (FII) and Fisher information weighted by a posterior distribution (FIP). Methods developed under the latter approach use Kullback-Leibler (KL) information, which is a global information measure (Chang & Ying, 1996; Chen, Ankenmann, & Chang, 2000).

Recent studies examining the efficacy of these 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. Further, the average test lengths under the alternative procedure were shorter than those under EAP estimation.

In their discussion, Chen, Ankenmann & Chang (2000; p. 253) suggest that “nothing is to be lost” by incorporating alternative item selection procedures in a CAT. Given the apparent convergence in performance among the more common FI item selection and the alternative item selection procedures after approximately 10 items, an interesting question is, “what is to be gained?” Answering this question requires a method for evaluating the inefficiency in CAT item selection, thereby suggesting how much room remains for improvement.

A related issue, not directly addressed in the current literature on CAT item selection, 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 is not directly influenced by errors in ability estimation.[1] Through this measure, a precise definition of efficiency may be obtained, one which follows naturally from the statistical concepts of efficiency and relative efficiency.

In order to utilize the concept of relative efficiency, it is useful to consider two tests, A and B, administered to an examinee possessing a 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 $I_A^{(T)}(\theta)/I_B^{(T)}(\theta)$. If test A is more efficient than test B (i.e., test A

[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 only of item parameters and a given value of ability.

yields more precise measurements at θ), $RE\left(A, B \middle| \theta\right) > 1$. Conversely, if test B is more efficient, $RE\left(A, B \middle| \theta\right) < 1$.

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* 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.

The argument presented here for a measure of efficiency parallels that of Spanos' (1999, p. 609) discussion of relative and full efficiency. He maintains that since relative efficiency alone is not necessarily useful (e.g., a poor estimator is relatively more efficient than an even poorer estimator), some fixed point of reference, namely the Cramer-Rao lower bound for the variance of an estimator, is required. In IRT applications where ability estimates are obtained by maximum likelihood (ML), the asymptotic variance of the ML estimator is in fact equal to the Cramer-Rao lower bound, or $1/I^{(T)}(\theta)$, where $I^{(T)}(\theta)$ is the accumulated test information at the true ability θ. However, the items contributing to $I^{(T)}(\theta)$ in a CAT environment are not fixed, and hence an additional lower bound—this one resulting from the choice of items—is needed to define efficiency. It is the quantity $I_0^{(T)}(\theta)$ described above that sets the lower bound on the variance for an estimator of θ (or, an upper bound on the precision of such an estimator) in the CAT environment. Thus, the ratio $I_{CAT}^{(T)}(\theta)/I_0^{(T)}(\theta)$ is not simply a measure of relative efficiency, but of efficiency itself, and may be used to define the efficiency of a CAT. The primary advantage of this definition is that the efficiency of item selection from different procedures (e.g., alternative item selection procedures or alternative ability estimation procedures) may be compared to a fixed point of reference, one which characterizes the most efficient estimator possible.

With a clearly defined measure of CAT efficiency, it is possible to identify any gaps in the efficiency of item selection. Thus, the question of what is to be gained by utilizing alternative procedures may be addressed directly, as the difference between the efficiency measure for a particular procedure and the upper bound on efficiency may be quantified. It is this difference that indicates exactly how much room for improvement exists.

Statement of the Problem

The efficiency of CAT item selection is dependent on item selection procedures as well as ability estimation procedures. Most commonly, maximum Fisher information (FI) item selection is employed in conjunction with either maximum likelihood (ML) or modal a posteriori (MAP) ability estimation. Because maximum FI item selection (under either ML or MAP) has been criticized as being inefficient, the first purpose of this study is to quantify the efficiency (or, inefficiency) of this most common item selection procedure. The second purpose of this study is to propose an alternative ability estimation procedure that addresses potential inefficiencies in CAT item selection, where this alternative procedure operates concurrently with either ML or MAP estimation and functions as an adjustment to either of these procedures. The third purpose of this study is to evaluate the efficiency of CAT item selection given five ability estimation procedures (i.e., ML, MAP, GSS, the proposed alternative procedure concurrent with ML, and the proposed alternative procedure concurrent with MAP) and two item selection procedures (i.e., maximum FI and maximum FII). Efficiency in this context is defined as above; namely, the ratio $I_{CAT}^{(T)}(\theta)/I_0^{(T)}(\theta)$. As this definition is predicated on evaluating information measures at an examinee's true ability θ, a simulation methodology is necessary, one which simulates a CAT administration for the following configurations: (a) item selection by maximum FI, and ability estimation either by ML, MAP, GSS, ML concurrent with the proposed alternative procedure, or MAP concurrent with the proposed alternative procedure; and (b) item selection by maximum FII, and ability estimation either by ML, MAP, GSS, ML concurrent with the proposed alternative procedure, or MAP concurrent with the proposed alternative procedure.

Maximum 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 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. 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. The item with the greatest value for the FII index is then selected for administration.

Ability estimation by both Xiao's (1999) 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 observed and expected scores given $\hat{\theta}_1$. 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; this alternative procedure is also based on a series of hypothesis tests. 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 $j+1^{\text{th}}$ item, while rejection of the null suggests a modified ability estimation procedure.

Two primary research questions thus follow:

1. How might the efficiency of maximum FI item selection under conventional ability estimation procedures be characterized, especially at the early stages of a CAT administration? More specific questions subsumed under this primary research question include: (a) After a fixed number of items have been administered, to what extent does efficiency of maximum FI item selection under ML or MAP ability estimation vary for different points along the ability continuum? For example, after *j* items have been administered, how does the efficiency of item selection at the middle of the ability distribution compare with efficiency in the tails? (b) What is the effect of ability estimation procedure on the efficiency of maximum FI item selection? For example, to what extent does efficiency vary depending on whether maximum likelihood (ML) estimation, a classical approach, or modal a posteriori (MAP) estimation, a Bayesian approach, is chosen?

2. Is it possible to improve upon the efficiency of maximum FI item selection under conventional ability estimation procedures by utilizing alternative item selection procedures, alternative ability estimation procedures, or a combination of both? This question is most relevant in cases where sizeable gaps in the efficiency of maximum FI item selection in combination with conventional ability estimation procedures (e.g., ML or MAP) have been identified. In parallel with the first primary research question, it is also appropriate to consider the following for this second primary research question: (a) After a fixed number of items have been administered, to what extent do the efficiency measures for the alternative item selection and ability estimation procedures vary for different points along the ability continuum? Specifically, how do the alternatives to FI

item selection with ML or MAP ability estimation compare to one another? (b) How do these efficiency measures compare with those obtained for maximum FI item selection with ML or MAP ability estimation? That is, to what extent are the alternative item selection and ability estimation procedures more (or less) efficient than maximum FI item selection in conjunction with conventional ability estimation procedures?

The two primary research questions were addressed using a simulation methodology. The CAT simulations employed here draw on an item bank of 367 pre-calibrated and dichotomously-scored 3P items from a recently-administered large-scale CAT assessment of mathematics ability. In the logistic metric where the scaling parameter $D = 1.7$, the mean and standard deviation of the discrimination parameters (i.e., a parameters) from the 367 items are 0.950 and 0.341, respectively. For the difficulty parameters (i.e., b parameters), the mean and standard deviation are 0.158 and 1.113, respectively. For the pseudo-guessing parameters (i.e., c parameters), the mean and standard deviation are 0.144 and 0.105, respectively. In its operational form, the CAT administered using this item bank is fixed at a length of 28 items; however, as it was hypothesized that the greatest variation in CAT efficiency would occur much earlier (e.g., at or before the 10th administered item), the CAT simulations were fixed such that no test exceeded a length of 25 items.

The four factors in the experimental design were: (1) item selection procedure (maximum FI or maximum FII item selection); (2) ability estimation procedure (ML, MAP, GSS, ML concurrent with the proposed alternative procedure, or MAP concurrent with the proposed alternative procedure); (3) true ability level at discrete points along the ability continuum (at -2, -1, 0, +1, or +2 logits); and (4) test length (5, 10, 15, or 25 items). For each of the experimental conditions, 1000 replications were generated. ML and MAP ability estimation procedures were included for the following reasons: (1) ML and MAP estimators behave differently, with the classical ML estimators being less biased than MAP estimators but prone to variability, while the Bayesian MAP estimators are biased but are less variable; (2) ability estimation by expected a posteriori (EAP), another popular Bayesian procedure, yields similar point estimates as MAP. The choice of the five discrete ability points is consistent with previous literature examining item selection procedures in CAT. Similarly, the CAT item selection literature suggests that alternative item selection procedures are most effective early in a CAT administration, with performance gains typically observed before the 10th administered item.

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 (that is, the range in the efficiency measure between the 25th and 75th percentile points) is reported as a measure of variability. Efficiency at both the 25th and 75th percentile points is also provided. In addition, the mean and standard deviation for the distribution of provisional ability estimates under each of the experimental conditions are reported.

Significance of the Study

The recognition that provisional ability estimates at the early stages of testing are prone to error has recently generated an area of research which seeks to improve the efficiency of a CAT primarily by means of alternative item selection procedures. The argument for these alternative item selection procedures is that maximum FI item selection, the most commonly used item selection procedure, is inefficient at the early stages of a CAT because it selects items whose information is at the maximum of an inaccurate or imprecise provisional ability estimate as opposed to an examinee's true ability.

What has not yet been fully considered is a method for comparing procedures against a common metric, such that the degree of efficiency (or inefficiency) in item selection is readily quantified. An upper bound for the efficiency of CAT item selection is proposed here; this upper bound makes possible an efficiency measure that is applicable regardless of the specific choice of item selection or ability estimation procedure.

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 addresses the question, to what *extent* is maximum FI item selection with these ability estimation procedures inefficient? As this study is able to assess the gaps in efficiency, it further addresses 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. If in fact maximum FI item selection with conventional ability estimation procedures is deemed inefficient under certain of the experimental conditions explored here, alternative item selection and ability estimation procedures that are relatively easy to implement in an operational setting are suggested and evaluated.

Limitations of the Study

The primary limitation of this study is that it is, by necessity, a simulation study. An examinee's true ability level θ must be known in advance in order to evaluate the efficiency measure $I_{CAT}^{(T)}(\theta)/I_0^{(T)}(\theta)$. Further, controlled comparisons among the item selection and ability estimation procedures considered here can only be made using a simulation design.

The items within the pool selected for this simulation study were calibrated from actual examinee response data and were utilized in an operational CAT administration for a large-scale assessment of mathematics ability. All items considered in this study are modeled under the 3-parameter logistic IRT model, and are thus dichotomously-scored. In addition, the alternative item selection procedure proposed here assumes that only dichotomously-scored items are being administered, although it should be possible to generalize this procedure for multiple-category IRT models such as the graded response model (Samejima, 1969). Thus, the study methodology is limited to dichotomously-scored items. Further, the results are dependent on the specific item pool selected for the study, and generalization to other CAT item pools should be approached with caution. It should be noted, however, that while the item response is indeed simulated in this design, the item parameters used for the simulation were calibrated from actual examinee response data.

The purpose of this study is to quantify the efficiency of two item selection procedures and five ability estimation procedures under what may be considered idealized or “best-case” scenarios. The rationale here is to minimize the influence of extraneous sources of variance on the measurement of CAT efficiency. It follows then, for example, that the simulations conducted for this study do not utilize any form of item exposure control. The addition of exposure control would serve only to lower the measured efficiency of the CAT item selection procedures considered here, which is at odds with the purpose of the study. Further, non-model-fitting responses or aberrant response patterns are not considered in this study.

CHAPTER 2

Review of Related Literature

A computerized adaptive test (CAT) is a form of tailored testing, as the goal of a CAT is to administer items to examinees that are well-targeted to their abilities. In addition to being a tailored test, a CAT is also a dynamic test, since it is assembled item-by-item based on the responses it observes from the examinee being tested. Test assembly in this adaptive framework relies on procedures for estimating examinee ability and selecting items for administration to that examinee. Both of these procedures must function in real-time, that is, they must operate concurrently with an examinee taking the test, as they are responsible for the test's construction.

In CAT, ability estimation procedures and item selection procedures are interdependent. Ability estimates serve to select the next item for administration, as the goal of a CAT is to administer items targeted to an examinee's ability. In turn, examinee responses to the administered items are used to estimate an examinee's ability. If the administered items are accurately targeted to an examinee's ability and the IRT model is appropriate, then greater precision in the measurement of that examinee's ability will be obtained. Once an examinee has completed the test, the ability estimation procedure is also responsible for obtaining a final ability estimate for that examinee, as well as the associated standard error associated with that estimate.

At present, a number of ability estimation and item selection procedures are available for use in a CAT. No one single ability estimation procedure or item selection procedure is generally agreed upon as being superior. How each compares among the others has been the focus of relatively recent research (e.g., most within the past 10-15 years, though insights by Birnbaum (1968) and Lord & Novick (1968) were voiced considerably earlier). The aim of this review is to provide a theoretical background for these procedures, to describe how each is utilized in the CAT environment, and to evaluate their performance in research studies.

Among the ability estimation procedures, five are considered here: maximum likelihood (ML), modal a posteriori (MAP), expected a posteriori (EAP), weighted likelihood estimation (WLE), and golden section search (GSS). Among the item selection procedures, five are considered here: maximum Fisher information (FI), Fisher interval information (FII), Fisher information with a posterior weight function (FIP), Kullback-Leibler (KL) information, and KL information weighted by a posterior density (KLP).

IRT ability estimation procedures in a computerized adaptive test

A computerized adaptive test (CAT) necessarily rests on procedures for estimating examinee ability, as a CAT attempts to adapt the delivery of administered items to the ability of the examinee being tested. In addition to provisional ability estimates—estimates of ability used to help select the items administered to an examinee—the CAT must also produce a final ability estimate, which is used to report an examinee's final score. Item response theory (IRT) provides a method for calculating both provisional and final ability estimates.

In the sections that follow, three aspects of IRT ability estimation in a CAT are explored: (1) the theoretical development of IRT ability estimation; (2) the implementation of ability estimation in an operational CAT; and (3) the evaluation of ability estimation procedures in terms of accuracy and precision of measurement.

Theoretical development

In IRT, it is assumed that person parameters (latent abilities) and item parameters underly examinee response behavior to an administered item. Response behavior in this sense is the probability of an examinee with latent ability θ providing a correct response to that item, or $P(U_i = 1|\theta)$ for item *i*. The relationship between θ and this response behavior is governed by the item characteristic curve, which is a function described by a set of parameters. In IRT, these parameters are called item parameters.

An examinee's true ability θ cannot be measured directly and so must be estimated. The estimation utilizes two sets of information: first, the pattern of responses observed from the examinee; and second, the item parameters from the items administered to the examinee. If the pattern of observed responses is denoted by $\mathbf{u} = \{u_1, u_2, \ldots, u_{N_I}\}$ and the item parameters by $\boldsymbol{\omega} = \{\omega_1, \omega_2, \ldots, \omega_{N_I}\}$, where N_I indicates the number of items, then the estimation of θ focuses on the probability $P(\theta|\mathbf{U} = \mathbf{u}, \boldsymbol{\omega})$. What is necessary in practical problems of estimation is a method for relating this posterior distribution $P(\theta|\mathbf{U} = \mathbf{u}, \boldsymbol{\omega})$ to the likelihood function, denoted by $P(\mathbf{U} = \mathbf{u}|\theta, \boldsymbol{\omega})$.

A solution to this problem is Bayes' Theorem (Hambleton & Swaminathan, 1985; Suen, 1990). Using this theorem, we find that

$$P(\theta|\mathbf{U}=\mathbf{u},\boldsymbol{\omega}) \propto P(\mathbf{U}=\mathbf{u}|\theta,\boldsymbol{\omega})P(\theta) \qquad \text{(Eq. 1)}$$

where $P(\theta|\mathbf{U}=\mathbf{u},\boldsymbol{\omega})$ is the posterior distribution of θ,
$P(\mathbf{U}=\mathbf{u}|\theta,\boldsymbol{\omega})$ is the likelihood function,
and $P(\theta)$ is the prior distribution of θ

IRT provides a mechanism for computing the likelihood function $P(\mathbf{U}=\mathbf{u}|\theta,\boldsymbol{\omega})$, as will be shown next. The choice of the prior distribution $P(\theta)$ will be discussed shortly.

The principle of local independence figures prominently in the calculation of IRT likelihood functions. Local independence requires that, for a fixed value of θ and a set of N_I test items, the joint distribution of $P(\mathbf{U}=\mathbf{u}|\theta,\boldsymbol{\omega})$ is equal to the product of the marginal probabilities $P(U_i = u_i|\theta,\omega_i)$ for items i = 1, 2, …, N_I (Lord and Novick, 1968). Thus,

$$P(U_1 = u_1, U_2 = u_2, \ldots, U_{N_I} = u_{N_I}|\theta,\boldsymbol{\omega}) = \prod_{i=1}^{N_I} P(U_i = u_i|\theta,\omega_i) \qquad \text{(Eq. 2)}$$

The assumption of local independence greatly simplifies the computation of the likelihood function and consequently the posterior distribution.

In addition to the likelihood function, Equation 1 requires a prior distribution $P(\theta)$ to be specified. The choice of the prior is often left to the researcher. If there is reasonable evidence to suggest a distributional form for the distribution of ability, then an informative prior may be used. (Note that this choice is Bayesian in nature.) If the researcher prefers to make no assumptions on the distribution of ability, then a non-informative prior (i.e., a uniform distribution) may be used. (Note that this choice is classical in nature.)

Once the posterior distribution is computed from the likelihood function and the prior distribution, ability estimation is possible. Four types of likelihood-based estimators are discussed here: maximum likelihood (ML), modal a posteriori (MAP), expected a posteriori (EAP), and Warm's weighted likelihood estimation (WLE). Note that the fifth ability estimator discussed here, GSS, does not utilize a likelihood function; rather, it compares optimally-weighted observed and expected scores under a golden section search optimization strategy.

In maximum likelihood estimation, the prior distribution $P(\theta)$ in Equation 1 is non-informative. Thus, both the maxima of the posterior distribution and the likelihood function occur at the same value along the ability scale. This value is the maximum likelihood estimate of an examinee's true ability θ, and is indicated by $\hat{\theta}_{ML}$. Rather than maximizing the likelihood function itself to identify $\hat{\theta}_{ML}$, the logarithm of the likelihood function is typically used as it simplifies calculations. If this log-likelihood is denoted by log L, then the maximum likelihood estimate $\hat{\theta}_{ML}$ is found by solving for the value of θ which satisfies the equation

$$\frac{\partial \log L}{\partial \theta} = 0 \qquad \text{(Eq. 3)}$$

Modal a posteriori (MAP) is similar to ML estimation in that the maximum of the posterior distribution is found to estimate θ. However, in MAP estimation an informative prior is chosen, often the normal density. The mode of the posterior distribution is then the estimate $\hat{\theta}_{MAP}$. If the expected value of the posterior distribution is found instead of the modal point, then the expected a posteriori (EAP) estimate $\hat{\theta}_{EAP}$ is obtained. Typically, Gauss-Hermite quadrature is used to find $\hat{\theta}_{EAP}$, such that

$$\hat{\theta}_{EAP} = \frac{\sum_{k=1}^{q} X_k L(X_k) W(X_k)}{\sum_{k=1}^{q} L(X_k) W(X_k)} \qquad \text{(Eq. 4)}$$

where X_k is one of q quadrature points, $W(X_k)$ is a weight associated with the quadrature point (e.g., corresponding to the normal density prior), and $L(X_k)$ is the likelihood function evaluated at X_k.

Although ML estimates are convenient due to their properties of asymptotic consistency and asymptotic normality, they are biased (Wang & Vispoel, 1998). Lord (1983) derived an expression for the bias in ML estimates, showing that they are biased outwards. Warm (1989) mentions that the magnitude of this bias is larger for negative values of $\hat{\theta}_{ML}$ than for positive values. Warm's weighted likelihood estimation (WLE) attempts to correct for this bias in ML estimates up to order n^{-1} by removing this first-order bias term from the ML estimates. The weighting function employed in Warm's procedure is specified in advance, but makes no assumptions about the distribution of θ. Rather, it is only a function of the items chosen for the test and so should not be confused with the informative prior distribution discussed earlier.

If the weighting function in WLE is denoted by $w(\theta)$, then the weighted likelihood estimate $\hat{\theta}_{WLE}$ is the solution to the equation (Warm, 1989)

$$\frac{\partial \log L}{\partial \theta} + \frac{\partial \ln w(\theta)}{\partial \theta} = 0 \qquad \text{(Eq. 5)}$$

The weighting function $w(\theta)$ is a function of test information, a concept to be discussed shortly.

When response patterns are either completely correct or completely incorrect, ML ability estimates cannot be determined, whereas estimates from MAP, EAP, and WLE are available. The problem with ML estimation in such cases lies in the fact that the likelihood function does not possess a maximum; hence, the limiting solution to Equation 3 is $\hat{\theta} \rightarrow -\infty$ for completely incorrect response patterns and $\hat{\theta} \rightarrow +\infty$ for completely correct patterns. In practice, such estimates are untenable; typically, examinees with such response patterns are either removed prior to ML estimation and assigned a score afterwards, or bounds are imposed on the ML estimation algorithm. It is common to choose ± 4 as the bounds for the ability scale.

While completely correct or incorrect response patterns may occur relatively infrequently for a fixed-length linear test, they are certainly guaranteed just after the first item is administered in a CAT and still remain likely early in a CAT administration. If an ability estimator with a non-informative prior (such as ML) is chosen for a CAT, it is desirable to force a variation in item responses as early as possible; otherwise, an ability estimate will not be available. This problem may be circumvented by using an estimator such as MAP or EAP.

Unlike ML, MAP, EAP, or WLE ability estimation, Xiao's (1999) golden section search (GSS) ability estimation procedure does not utilize a likelihood function to obtain an ability estimate. Rather, it uses an optimization strategy to search the ability continuum for an ability estimate consistent with the observed pattern of responses. The first search interval covers the entire (bounded) range of the ability continuum, typically

[-3, +3] or [-4, +4]. If this first search interval is denoted by [a, b], then the midpoint c of this interval is taken as the first possible estimate of ability. A hypothesis test is conducted at this point, where the optimally-weighted observed and expected scores, which are functions of the observed and expected proportions, are compared. If the null hypothesis is not rejected, then this midpoint c is accepted as the provisional ability estimate and no further search is executed. On the other hand, if the null hypothesis is rejected, then a new search section $[a', b']$ is identified, such that [a, b] is reduced by the golden section ratio to obtain the new $[a', b']$. If the sign of the test statistic is negative (i.e., observed score is less than the expected score), then the original section is reduced by the golden section ratio so that $a' = a$ and $b' = a + t(b - a)$, where t is equal to the golden ratio[2] $\left(\sqrt{5} - 1\right)/2$. If the sign of the test statistic is positive, then $b' = b$ and $a' = b - t(b - a)$. The midpoint c' of this new section is then taken as the next possible estimate of ability, and another hypothesis test is conducted. This process of hypothesis-testing followed by interval sectioning continues as long as the null hypothesis is rejected. Once a failure to reject the null hypothesis is reached, the midpoint of the section is taken as the provisional ability estimate $\hat{\theta}_{GSS}$. Xiao (1999) considers this estimate to be equivalent to the maximum likelihood estimate $\hat{\theta}_{ML}$.

Xiao uses optimal scoring weights (Birnbaum, 1968) to arrive at observed and expected scores. For any item i, the optimal scoring weight is given by

[2] The golden ratio may be found by construction: given a line divided into two segments a and b with $a < b$, the ratio of a to b should be equal to the ratio of b to the entire line length $a + b$. Let t be the golden ratio relating segments a and b, such that $a = tb$. Then by construction $tb/b = b/(tb + b)$, which reduces to the quadratic equation $t^2 + t - 1 = 0$. The quadratic equation is satisfied by $t = \left(-1 \pm \sqrt{5}\right)/2$; the positive root is taken here.

$$W_i(\theta) = \frac{P_i'(\theta)}{P_i(\theta)[1 - P_i(\theta)]} \qquad \text{(Eq. 6)}$$

Thus, an optimally-weighted observed score for an examinee with responses u_i may be computed as

$$X = \sum_{i=1}^{n} W_i u_i \qquad \text{(Eq. 7)}$$

The optimally-weighted expected score for *X*, given ability θ is computed as

$$E[X|\theta] = \sum_{i=1}^{n} W_i P_i(\theta) \qquad \text{(Eq. 8)}$$

and the variance by

$$Var[X|\theta] = \sum_{i=1}^{n} W_i^2 P_i(\theta)[1 - P_i(\theta)] \qquad \text{(Eq. 9)}$$

Then the test statistic employed by Xiao is

$$z = \frac{X - E[X|\theta]}{\sqrt{Var[X|\theta]}} \qquad \text{(Eq. 10)}$$

Rejection of the null hypothesis occurs when the absolute value of test statistic *z* exceeds a critical value z_c. Xiao (1999) recommends that a value $z_c = 0.7$ be used, based on prior studies examining the golden section search strategy. Further, in cases where the optimally weighted scores given by Equation 7 are less than

$$X_{\min} = \sum_{i=1}^{n} W_i c_i \qquad \text{(Eq. 11)}$$

where the c_i are the pseudo-guessing parameters from the 3P IRT model, the optimally weighted observed scores *X* are replaced by $X_{\min}$.

Estimating standard errors

In addition to finding a point estimate for θ, it is often desirable to estimate standard errors of measurement ($SE_{\hat{\theta}}$). IRT provides a method for computing conditional standard errors of measurement, thus characterizing the varying precision of a test along the ability scale. For all procedures except EAP, $SE_{\hat{\theta}}$ is found by means of the test information function[3]. The test information function $I(\theta)$ is defined as

$$I(\theta) = -E\left[\frac{\partial^2 \log L}{\partial \theta^2}\right] \tag{Eq. 12}$$

The asymptotic variance of the ML estimate $\hat{\theta}_{ML}$ is then related to the test information function $I(\theta)$ by (Hambleton & Swaminathan, 1985)

$$Var\left(\hat{\theta}_{ML} | \theta\right) = \left[I(\theta)\right]^{-1} \tag{Eq. 13}$$

and hence $SE_{\hat{\theta}_{ML}} = \left[I(\theta)\right]^{-\frac{1}{2}}$. Further, it is known that $\hat{\theta}_{ML}$ is asymptotically normally distributed. For WLE, Warm (1989) notes that the estimate $\hat{\theta}_{WLE}$ maintains similarities with the ML estimator, in that $\hat{\theta}_{WLE}$ is asymptotically normally distributed, with asymptotic variance equal to the asymptotic variance of $\hat{\theta}_{ML}$. Thus, $SE_{\hat{\theta}_{WLE}} \cong SE_{\hat{\theta}_{ML}}$.

[3] Xiao (1999) does not explicitly provide the standard error associated with the GSS point estimate; however, as the GSS estimator is assumed to be equivalent to the ML estimator, it is likely that the assumed equivalence carries over to the standard errors as well.

The standard error of measurement $SE_{\hat{\theta}_{MAP}}$ for the Bayes MAP estimate is usually smaller than that obtained for the ML estimate, and is given by (Wainer & Mislevy[4], 1990)

$$Var\left(\hat{\theta}_{MAP}\middle|\theta\right)=\left[I(\theta)-E\left(\frac{\partial^2 \ln p(\theta)}{\partial\theta^2}\right)\right]^{-1} \qquad \text{(Eq. 14)}$$

where $p(\theta)$ is the prior distribution of θ. For a prior distributed as $N(0,1)$, the term $E\left(\frac{\partial^2 \ln p(\theta)}{\partial\theta^2}\right)=-1$.

The standard error of measurement $SE_{\hat{\theta}_{EAP}}$ for the EAP estimate is computed from the posterior distribution instead of from the test information function. As shown by Wang & Vispoel (1998), the variance of the posterior distribution $g(\theta)$ is

$$Var\left(\theta\middle|u\right)=\int_{-\infty}^{\infty}\theta^2 g\left(\theta\middle|u\right)d\theta-\left(E\left(\theta\middle|u\right)\right)^2 \qquad \text{(Eq. 15)}$$

Gauss-Hermite quadrature may be used to approximate this integration. The resulting expression for the estimated variance is then

$$\hat{\sigma}^2\left(\hat{\theta}_{EAP}\right)=Var\left(\theta\middle|u\right)=\frac{\sum_{k=1}^{q}\left(X_k-\hat{\theta}_{EAP}\right)^2 L\left(X_k\right)W\left(X_k\right)}{\sum_{k=1}^{q}L\left(X_k\right)W\left(X_k\right)} \qquad \text{(Eq. 16)}$$

where X_k is one of q quadrature points, $W(X_k)$ is a weight associated with the quadrature point (e.g., corresponding to the normal density prior), and $L(X_k)$ is the likelihood function evaluated at X_k. From Equation 16, the standard error $SE_{\hat{\theta}_{EAP}}=\sqrt{\hat{\sigma}^2\left(\hat{\theta}_{EAP}\right)}$.

[4] Their equation, as printed in the text, has been corrected here.

Implementation of ability estimation in CAT

Unlike a conventional fixed-length linear test, where ability estimation occurs after the examinee has completed the entire test, a CAT estimates an examinee's ability in a sequential fashion. The CAT will estimate an examinee's ability after each item has been administered; thus, for any given examinee, a CAT will have as many ability estimates for that examinee as the number of items administered to that examinee.

Some nomenclature is useful for identifying the ability estimates obtained in a CAT. Suppose that N_I items are administered to an examinee by a CAT, and for each of these items an ability estimate $\hat{\theta}_i$ where $i = \{1, 2, \ldots, N_I\}$ is available. Then the estimates $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_{N_{I-1}}$ are called the *provisional ability estimates* and the estimate $\hat{\theta}_{N_I}$ is called the *final ability estimate*. The provisional ability estimates are used by the CAT to select the next item for the examinee; the final ability estimate is taken as the best estimate of that examinee's ability and is used to report a final score on the test.

Calculation of provisional ability estimates during a CAT administration occurs each time an examinee responds to an administered item. At the start of a CAT, typically no information is available about an examinee's ability since no items have been administered. However, adaptive tests could conceivably use collateral information to make a very general classification of an examinee's proficiency. After the examinee has responded to the first item, an estimate of ability may be attempted. If the item parameters for this first item ($i = 1$) are denoted by ω_1 and the examinee response by u_1, then an estimate $\hat{\theta}_1$ may be calculated using Equation 1 and one of the four ability estimation procedures described here. However, if ML estimation is used, no ability estimate will be identified since the likelihood function will not possess a maximum.

Boundary conditions, as described above, may be imposed. For example, if $u_1 = 1$, then $\hat{\theta}_{ML} | u_1, \omega_1 = +4$. If either MAP, EAP, or WLE are used, a bounded ability estimate will be obtained. In addition, if GSS is used, a bounded ability estimate is always guaranteed, as GSS is a search strategy employed over a bounded interval.

An item selection algorithm will choose the second item for the examinee based on the first provisional estimate $\hat{\theta}_1$. After responding to this second item, the examinee's response vector is $\mathbf{u} = \{u_1, u_2\}$ and the vector of item parameters $\boldsymbol{\omega} = \{\omega_1, \omega_2\}$. The second provisional ability estimate, $\hat{\theta}_2$, may then be calculated from the available information, namely $\mathbf{u}$ and $\boldsymbol{\omega}$. Note that if $u_1 = u_2$, then an ML estimate will still be unidentified unless boundary conditions are imposed.

As the CAT administration proceeds, provisional ability estimates are updated with each successive item. After the last item is administered, the ability estimate $\hat{\theta}_{N_I}$ is computed for the vector of responses $\mathbf{u} = \{u_1, u_2, \ldots, u_{N_I}\}$ and item parameters $\boldsymbol{\omega} = \{\omega_1, \omega_2, \ldots, \omega_{N_I}\}$. This $\hat{\theta}_{N_I}$ is taken as the final ability estimate for the examinee. Notice that if any ability estimation procedure is used, this final ability estimate will be identical to the ability estimate obtained for a fixed-length linear test possessing the same set of items $i = \{1, 2, \ldots, N_I\}$. This equivalence is due to the estimation procedures' indifference towards the order in which items are presented. ML, MAP, EAP, and WLE use the likelihood function shown in Equation 2; in that expression, the order of item responses is not taken into account. GSS uses optimally weighted observed and expected scores; again, the order of item responses does not influence the computation of these scores.

Although the estimation procedures themselves do not take order of item administration into account, plotting the sequence of provisional ability estimates against item administration number can be informative, and illustrates some features of ability estimation in CAT. In Figure 1, a plot of ML, MAP, and EAP provisional ability estimates versus item administration number for a particular examinee is shown. The source of this data is from a recent CAT administration of a large-scale assessment of mathematics ability; this test is a fixed-length CAT with $N_I = 28$.

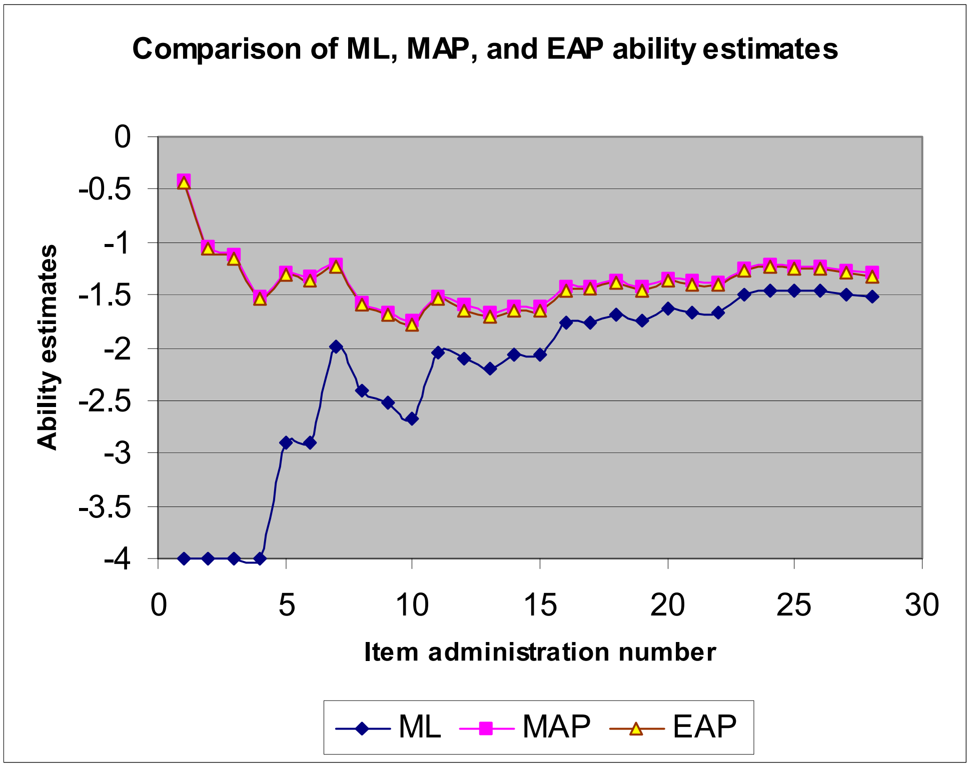

Figure 1. Comparison of ML, MAP, and EAP provisional ability estimates for an examinee with final ability estimates $\hat{\theta}_{ML}$ = -1.53, $\hat{\theta}_{MAP}$ = -1.30, and $\hat{\theta}_{EAP}$ = -1.33.

The first feature to note from these figures is that all three estimation procedures converge to a final ability estimate as the test progresses. From Figure 1, it can be seen that although the three procedures yield slightly different final ability estimates, they are rather close to one another. However, the trend in the sequence of provisional ability estimates from commencement to termination of testing is quite different between the ML procedure and the MAP and EAP procedures. Notice that the trend for ML estimates is upward, from a negative bound of -4 to the final estimate of -1.53, while the trend for the MAP and EAP estimates is downward, from the first provisional estimate of approximately -0.4 to the final estimate of approximately -1.3. Further, notice that the first four provisional estimates for ML are all at the negative bound of -4, whereas the corresponding estimates for MAP and EAP are not all equal. Finally, the variability in the ML estimates over time is greater than that of the MAP and EAP estimates. It should also be noted that the standard errors associated with the MAP and EAP estimates will always be smaller than those associated with the ML estimates.

The differences in the trends between the ML procedure and the MAP and EAP procedures may be attributed to the presence or absence of an informative prior distribution in estimating ability. The informative prior used in the MAP and EAP procedures reduces the variability in provisional ability estimates considerably. Further, it moves the initial estimates towards the mode of the prior; in this case, the prior is distributed as $N(0,1)$. Because this examinee incorrectly answered the first four items, ML estimation must impose a lower bound of -4 on the ability estimates, as a maximum in the likelihood function is not present even after four items have been administered. In contrast, ability estimates are possible in the MAP and EAP procedures after these four items because of the informative prior.

A comparison of the last few provisional ability estimates and the final ability estimates across the three procedures shows that they are similar, with final ability estimates $\widehat{\theta}_{ML} = -1.53$, $\widehat{\theta}_{MAP} = -1.30$, and $\widehat{\theta}_{EAP} = -1.33$. Although they are similar, they are not identical, and the slight difference is worthwhile to note. Even at the end of the test, the influence of the prior distribution remains; the two Bayesian procedures MAP and EAP yield estimates that are biased inwards towards the mode of the prior distribution as compared to the maximum likelihood estimate.

It was noted that the final estimate of ability is computed after the last item is administered in a CAT. How the CAT determines when to terminate testing depends on a decision rule, often referred to as a stopping criterion. Unlike a conventional fixed-length linear test, where testing terminates after a fixed number of items have been administered, CAT has no such restriction. In principle, two subsets of items administered by a CAT to two different examinees may be of different sizes; that is, the number $N_{I,1}$ of items administered to one examinee may be larger than, equal to, or smaller than the number $N_{I,2}$ of items administered to a different examinee. Further, the degree of overlap or number of items in common between the two subsets may vary from 0, 1, …, $\min(N_{I,1}, N_{I,2})$.

Thissen & Mislevy (1990) describe three CAT stopping criteria: (1) terminate testing after a specific number of items have been administered; (2) terminate testing when a specific precision in measurement has been reached (e.g., a minimum standard error of measurement for all examinees); and (3) terminate testing after a specific amount of time has elapsed. They note that any one of these criteria may be chosen for a CAT, or a mixture of these criteria may be employed. For power tests, however, the first two criteria are most relevant and the advantages and disadvantages of these criteria are discussed below.

Advantages for the first criterion, terminating testing after a specific number of items have been administered, include ease of implementation in a CAT algorithm and better prediction of item usage for items in the pool. The primary disadvantage of this criterion is that examinees will be measured with differing degrees of precision along the ability continuum. This disadvantage is of greatest concern for those examinees at the extremes of the ability scale; it is likely that the first few items selected for these examinees will not provide much information for estimating their abilities.

Advantages for the second criterion, terminating testing when a specific precision in measurement has been reached, include the assurance that all examinees will be measured with the same degree of precision. Where this property is particularly useful is in statistical analyses on the test data that make assumptions about the homogeneity of error variances. The disadvantage of this criterion is that by design, tests are of variable lengths. Predicting how much time is required to complete tests most likely involves simulating the CAT administration for examinees at different points along the ability continuum. Further, operational considerations such as the amount of “seat time” required for examinees to complete the test become more complex with variable length tests. This consideration is relevant for large-scale testing operations where the use of test center facilities is charged on a per unit time basis.

Evaluation of ability estimation procedures

Five ability estimation procedures have been discussed here: ML, MAP, EAP, WLE, and GSS. In order to evaluate how each performs in a CAT environment, a simulation methodology is required, since an examinee's true ability is never known in advance and so actual test data cannot provide an indication of how well ability is being estimated. To quantify the degree to which ability is being appropriately measured, simulation studies typically report one or more of the following: the bias of an ability estimate $\hat{\theta}$ from its true value θ, the standard error *SE*($\hat{\theta}$) associated with the ability estimate, and the root-mean-square error *RMSE*($\hat{\theta}$). The relationship among these three quantities is $RMSE^2 = Bias^2 + SE^2$.

Currently, there is no one single published article which compares ML, MAP, EAP, WLE, and GSS in a CAT environment. However, Wang & Vispoel (1998) compare the ML, MAP, and EAP estimation procedures, and Cheng & Liou (2000) compare the ML and WLE estimation procedures. Xiao (1999) compares the GSS and EAP procedures, but in a CAT environment that differs from Wang & Vispoel's (1998) and Cheng & Liou's (2000) studies. That is, Xiao (1999) examines a computerized adaptive classification test, whereas Wang & Vispoel (1998) and Cheng & Liou (2000) examine a CAT which provides final ability estimates for examinees, as opposed to one that assigns them to discrete proficiency categories. Although Warm (1989) compares the ML, MAP, and WLE procedures, the conditions under which the simulation study were conducted do not parallel those used in current adaptive testing practice; specifically, his study did not draw items from a pre-determined item pool, but rather generated item parameters during the simulation based on initial constraints (e.g., all discrimination parameters set to 2.0) and optimal item selection (e.g., item parameters corresponding to maximum information were generated as needed). Therefore, Warm's (1989) study utilized infinite item pools, whereas Wang & Vispoel (1998) and Cheng & Liou (2000) used finite item pools whose parameters were chosen a priori. Thus, the studies by Wang & Vispoel (1998), Cheng & Liou (2000), and Xiao (1999) will be discussed here. However, the discussion will focus primarily on Wang & Vispoel's (1998) and Cheng & Liou's (2000) studies, as the nature of their studies and their outcome measures are comparable.

Wang & Vispoel (1998) compared the ML, EAP, and MAP estimation procedures under a number of CAT administration conditions: (a) fixed- vs. variable-length CATs; (b) ideal vs. realistic item pools; and (c) effect of prior distribution on final ability estimates for EAP and MAP estimation. The first of these conditions compared ability estimates from fixed-length CATs (with items varying in number from 10 to 50 in steps of 10) to those from variable-length CATs (with target reliabilities of 0.80 and 0.90). The second of these conditions compared estimates across three types of item pools all of size $N_{pool} = 300$: (i) two ideal item pools with difficulty parameters b_i distributed uniformly over an interval [-3.6, 3.6], pseudo-guessing parameters c_i fixed at 0.15, and average discrimination parameters a_i drawn from a $N(1.1, 0.1)$ distribution for the moderate discriminating items, or from a $N(1.9, 0.1)$ distribution for the high discriminating items; and (ii) a realistic item pool with item parameters based on vocabulary items from six paper-and-pencil forms of the Iowa Tests of Educational Development. For the realistic item pool, average discrimination parameters were moderate and the majority of items were of middle difficulty. Finally, the third of these conditions used either a fixed prior distribution (i.e., $N(0,1)$) for all examines, or a variable prior distribution depending on the true ability of the examinee. In the variable prior conditions, the mode of the prior was chosen so as to be closest to an examinee's true ability; one of three possible prior distributions (with modes of -2, 0, or +2) was selected for an examinee.

In terms of the standard error, or $SE(\hat{\theta})$, Wang & Vispoel (1998) found that ML estimates overall had the highest standard errors, and that these were underestimated by the test information functions (see Equations 12 and 13) used to calculate the standard errors for these estimates. The standard errors for ML estimates were greatest at the extremes of the ability distribution, most likely due to the relative lack of items with difficulty parameters located at these points. For procedures such as EAP, where standard errors are based on the posterior distribution and not on the test information function, the estimated $SE(\hat{\theta})$ were consistent with those observed empirically in the simulation study. The manipulation of prior distribution type (fixed or variable) had little effect on $SE(\hat{\theta})$ for the MAP or EAP procedures.

While the Bayesian procedures performed well in terms of $SE(\hat{\theta})$, the non-Bayesian ML estimates were the least biased. When using the fixed prior distribution for the MAP and EAP procedures, ability estimates were biased inwards towards the mode of the prior. This bias was particularly strong for examinees at the extremes of the ability continuum. This bias in Bayesian estimates was exacerbated when few items were available at these extremes (i.e., under the realistic item pool condition); it should be noted, however, that the ML estimates were also biased at these extreme points, but only slightly so. Additionally, the bias in ML estimates at these extremes was outward, rather than inward. Of the Bayesian procedures, EAP showed less bias as compared to MAP.

However, if a variable selection of prior distribution was employed, the bias in MAP and EAP estimates was reduced, and quite significantly for the ideal item set conditions.

The focus of Cheng & Liou's (2000) study was twofold: first, it investigated the performance of ML and WLE estimation procedures in a CAT; and second, it investigated the performance of local and global information item selection procedures. For the purposes of this discussion, only the results pertaining to the ability estimation procedures will be included. Their study utilized 204 items from a 1992 NAEP Reading Assessment, where the range in discrimination parameters was 0.452 to 2.502, the range in difficulty parameters was -2.325 to 3.061, and the range in pseudo-guessing parameters was 0 to 0.373. The mean of the item parameters was $\overline{a}_i = 1.194$, $\overline{b}_i = -0.024$, and $\overline{c}_i = 0.124$. A fixed-length termination rule of 30 items was chosen for the simulation study.

The major finding from Cheng & Liou's (2000) study regarding ML and WLE estimation procedures was that WLE outperformed ML estimation in terms of less bias at the early stages of testing. However, after approximately 10 items were administered, both ML and WLE estimates generally possessed the same degree of bias. This result was anticipated by Wang & Vispoel (1998) when they discussed the bias function for the ML estimator, derived by Lord (1983). As items are targeted to examinee ability, the bias in $\hat{\theta}_{ML}$ approaches zero. Since Warm's WLE corrects for this bias, it is logical that as the bias approaches zero, the ML estimate will equal the WLE estimate. In fact, this convergence appears to be taking place in Cheng & Liou's (2000) study. At the early stages of testing, items are not well-targeted to examinee ability; thus, the ML estimates are biased and the WLE estimates, which have been corrected for this bias, were indeed found to be less biased than the ML estimates. However, as the CAT proceeds, and items are more accurately targeted to examinee ability, the WLE and ML estimates become similar.

If conclusions are to be drawn based on Wang & Vispoel's (1998) and Cheng & Liou's (2000) studies, it seems that the choice of ability estimation procedure must take into account the researcher's tolerance for bias versus measurement error. In general, the Bayesian procedures provide smaller measurement error at the expense of larger bias; the reverse is true for the non-Bayesian ML and WLE procedures. Although the degree of bias in the Bayesian approaches may be lessened by employing a variable selection of the prior distribution, this method rests on the assumption that such a distribution can be accurately identified for each examinee. Routing tests may provide one possible solution, but they themselves possess measurement error and could erroneously assign a prior distribution which might impose an even larger bias on ability estimates.

Xiao (1999) explored the EAP and GSS ability estimation procedures in a computerized adaptive classification test, or grading test. The outcome measures in Xiao's (1999) study differ from those in Wang & Vispoel's (1998) and Cheng & Liou's (2000) studies, and include: (a) the proportion of correct grade assignment decisions; and (b) average test length. Thus, Xiao's (1999) study employed variable-length CATs, as opposed to the fixed-length CATs of Wang & Vispoel (1998) and Cheng & Liou (2000).

Xiao (1999) simulated computerized adaptive grading tests where the item pool consisted of 200 pre-algebra items from an ACT Mathematics test. The range in discrimination parameters was 0.472 to 1.756, the range in difficulty parameters was −1.588 to 1.656, and the range in pseudo-guessing parameters was 0.056 to 0.494. The mean of the item parameters was $\overline{a}_i$ = 1.025, $\overline{b}_i$ = 0.086, and $\overline{c}_i$ = 0.172, and the standard deviations were 0.242, 0.666, and 0.067, respectively. Three specific grading tests were examined: the first possessed three proficiency levels; the second possessed five levels; and the third possessed 10 levels. Although each of the grading tests was variable in length, a maximum test length was specified for each. For the 3-, 5-, and 10-level tests, the maximum test length was 8, 20, and 20 items, respectively. Examinee ability was simulated by drawing from a *N*(0,1) distribution. The item selection method used for all conditions was maximum FI selection. For each test, 500 examinees were simulated; however, 10 tests were administered for each ability drawn from the distribution. Thus, the total number of replications per cell in her design was 5000.

Xiao (1999) found that there was “no general difference among the [ability estimation procedures] in the proportion of correct classifications” (p. 145). She notes that EAP estimation led to more correct classifications in the 3-level test; however, that effect was not observed in the 5- or 10-level tests. For the 5-level tests, approximately 79% of the classifications were correct; for the 10-level tests, approximately 58% of the classifications were correct. In the case of the 3-level tests, the proportion of correct classifications for GSS was 83.2%, while it was slightly higher at 85% for EAP. Thus, the GSS and EAP ability estimation procedures led to similar classification accuracy across the grading tests, with EAP being more accurate for the 3-level test.

With respect to average test length, a significant effect was observed for ability estimation procedure, with GSS resulting in slightly shorter tests than EAP in the case of the 3- and 5-level tests. For the 10-level tests, the stopping rule could not be met before the maximum number of items (i.e., 20) was reached; thus, no effect of ability estimation procedure was observed for the 10-level tests. For the 5-level test, GSS led to an average test length of 7.36 items, while the average test length for EAP was 7.50. For the 10-level test, the average test length for GSS was 19.68, whereas it was 19.90 for EAP. Overall, it appears that GSS is as effective as EAP in classifying examinees in a grading test, and that some reduction in average test length occurs for the GSS ability estimation procedure.

Item selection procedures in a computerized adaptive test

In traditional test design, test developers carefully select items for a test form based on content specifications and psychometric properties. Drafts of these test forms are reviewed by test developers for quality control; once approved, the forms are ready for distribution. In a computerized adaptive test (CAT), however, test assembly is altogether different. Tests are dynamic, in that they are constructed during the administration of the test itself. Because the goal of a CAT is to tailor the administered test to examinee ability, tests cannot be assembled in advance. Rather, the test must be assembled item-by-item in real time.

In the sections that follow, four aspects of item selection procedures in a computerized adaptive test are explored: (1) an overview of computerized adaptive test assembly; (2) some item selection procedures available for CAT; (3) operational constraints imposed on CAT item selection, including content balancing, item exposure, and locally-dependent items; and (4) an evaluation of item selection procedures.

Overview of CAT assembly

The goal of CAT is to administer items as closely matched to an examinee’s ability as possible. That is, administered items should be neither too easy nor too difficult for any given examinee. One advantage of such an adaptive, or tailored, test is that it can substantially reduce the number of items needed to achieve a desired level of measurement precision; often a test may be shortened to one-half its original length (Stocking, Smith & Swanson, 2000).

Clearly an adaptive test must utilize a mechanism for administering appropriate items to examinees. In a CAT, all items are drawn from one source, commonly called the item pool. This item pool contains all items which have been approved for administration to examinees; however, any given examinee should be administered only a subset of these items, and these items should be targeted as closely as possible to an examinee's ability. Within the item pool is information describing the item response theory (IRT) parameters for each item, such as item difficulty, discrimination, and pseudo-guessing parameters.

Item selection algorithms certainly take into account the psychometric properties of items in order to determine which items are most appropriate for examinees. However, there are a number of other constraints in item selection that must be considered. First, like any conventional test which is modeled after a blueprint, content specifications must be followed. Second, and what has become a challenging problem in CAT, is item exposure. Because CATs are administered continuously as opposed to periodically, security issues such as item exposure become a concern (Stocking & Swanson, 1996). Thus, administering a CAT is not simply a matter of choosing the optimal item based on psychometric criteria, blueprint specifications, or minimal item exposure, but a combination of all three constraints. Further, maximizing one or more of these constraints may come at the expense of minimizing any of the remaining constraints. These issues will be explored in greater depth in the third section, operational constraints in item selection.

Any CAT involves an interaction between item selection and ability estimation. Items are selected based on provisional ability estimates, and provisional ability estimates are calculated based on examinee responses to selected items. Although items could be selected based on the match between provisional ability estimates and item difficulty, it is more common for items to be selected based on the match between provisional ability estimates and item information content. These information selection procedures may further be classified into two categories: those that employ local information, and those that employ global information. These procedures are described in more detail in the next section.

A special case of item selection occurs at the very start of the test, when a provisional ability estimate is not available. Thus, at test commencement, the item selection procedure must be modified. Thissen & Mislevy (1990) point out that although no information may be available for a particular examinee at the start of a test, a great deal of information may exist regarding the distribution of abilities of examinees who have previously taken the test. In such a situation, then, the mean of this distribution may be taken as the initial estimate of an examinee's ability.

Item selection procedures

Citing Davey & Parshall (1995), Stocking & Lewis (1995) maintain that one of the goals in CAT is "to maximize test efficiency by selecting the most appropriate items for a test taker" (p. 4). From a psychometric perspective, it may not be immediately clear what criteria constitute "most appropriate". Thissen & Mislevy (1990) note that early in the development of computerized adaptive testing, it was suggested by Urry (1970) that the optimal item for selection was the one whose difficulty parameter (i.e., *b* value) was closest to an examinee's provisional ability estimate. They further note that Lord (1977) suggested a similar procedure.

Currently items are selected not based on their difficulty parameters, but on some measure of their information. Item information in IRT describes the precision with which an item can measure an examinee's ability. Because item information varies as a function of ability, any given item will be more informative at certain points along the ability continuum than at others. For a dichotomous item *i*, this item information function is given by

$$I_i(\theta) = \frac{\left[P_i'(U_i = 1|\theta)\right]^2}{P_i(U_i = 1|\theta)\left[1 - P_i(U_i = 1|\theta)\right]} \qquad \text{(Eq. 17)}$$

where $I_i(\theta)$ is the information provided by the item *i* at ability θ, $P_i(U_i = 1|\theta)$ is the probability of a correct response on item *i* given ability θ, and $P_i'(U_i = 1|\theta)$ is the first derivative $\frac{\partial}{\partial\theta}P_i(U_i = 1|\theta)$.

To describe the precision of an entire test, rather than a single item, the test information function is used. Conveniently, the test information function is the sum of the item information functions given in Equation 17; thus,

$$I_{1,2,\ldots N}(\theta) = \sum_{i=1}^{N} \frac{\left[P_i'(U_i = 1|\theta)\right]^2}{P_i(U_i = 1|\theta)\left[1 - P_i(U_i = 1|\theta)\right]} \qquad \text{(Eq. 18)}$$

where $I_{1,2,\ldots N}(\theta)$ is the test information function for a test consisting of *N* items. As noted by van der Linden & Pashley (2000), Birnbaum (1968) suggested that the test information function be used as the criterion for fixed-length linear test assembly. If maximum likelihood (ML) estimation is used to estimate ability, then the asymptotic variance of the ML estimator is the reciprocal of Equation 18.

van der Linden & Pashley (2000) note that while "no asymptotic motivation existed" for using information as the item selection criterion (since the number of items administered in a CAT would likely be well below the number required for asymptotic results), "the maximum-information criterion was immediately adopted as a popular choice" (p. 9). This method of maximum information item selection (or maximum Fisher information item selection) is employed as follows: given a provisional ability estimate $\hat{\theta}$ for an examinee, select an item j for administration to that examinee where $I_j\left(\hat{\theta}\right)$ is highest at that $\hat{\theta}$ among all other available items in the pool.

This use of Fisher information for item selection is perhaps the most popular. It is relatively straightforward to apply, and has the additional advantage that items may be sorted beforehand in what is called an information table (Thissen & Mislevy, 1990). This information table lists items in order based on information; these lists are further subdivided according to ability quadrature points. Quadrature points are discrete points along the ability scale that are used to approximate the ability continuum; they are also used in EAP ability estimation and typically number between 20 and 30 for a scale ranging from -4 to +4. Thus, once the CAT has obtained a provisional ability estimate $\hat{\theta}$ for an examinee, the item selection algorithm searches the information table for the entry whose quadrature point is closest to $\hat{\theta}$, and chooses the highest ranking item available for that entry. Since items have been pre-sorted according to an earlier calculation of item information (e.g., using Equation 17), no additional information calculations are necessary during the administration of the test. This method increases the computational efficiency of a CAT.

Nevertheless, variations on Fisher information item selection (FI) have been proposed. The argument for these methods is that FI does not take into account the error associated with the provisional ability estimate $\hat{\theta}$, which may be important to consider especially at the start of the CAT. Chen, Ankenmann, & Chang (2000) describe two variations, Fisher interval information (FII) and Fisher information with a posterior weight function (FIP), each based on a general weighted information criterion proposed by Veerkamp & Berger (1997). This general weighted information criterion (GWIC) is defined for an item *j* as

$$GWIC_j(\theta) = \int_{-\infty}^{\infty} W(\theta) I_j(\theta) d\theta \qquad \text{(Eq. 19)}$$

where $W(\theta)$ is a weight function and $I_j(\theta)$ is the information function for item *j*. In FII, it is proposed that the weight function be uniform over an interval defined in terms of the expected error associated with the provisional ability estimate $\hat{\theta}$, such that

$$W(\theta) = \begin{cases} 1, & \theta \in \left(\hat{\theta}_l, \hat{\theta}_u\right) \\ 0, & \text{otherwise} \end{cases} \qquad \text{(Eq. 20)}$$

with interval

$$\left(\hat{\theta}_l, \hat{\theta}_u\right) = \left(\hat{\theta} - \frac{z}{\sqrt{I_{1,2,\ldots,j-1}\left(\hat{\theta}\right)}}, \hat{\theta} + \frac{z}{\sqrt{I_{1,2,\ldots,j-1}\left(\hat{\theta}\right)}}\right) \qquad \text{(Eq. 21)}$$

where $I_{1,2,\ldots,j-1}\left(\hat{\theta}\right)$ is the test information for the items 1, 2, …, *j*-1 already administered, $\hat{\theta}$ is the provisional ability estimate after *j*-1 items, and *z* is a standard normal deviate specified for a desired degree of confidence. In contrast to FI, where an item is selected based on maximum information $I_j\left(\hat{\theta}\right)$ at a point $\hat{\theta}$, FII selects an item based on the maximum area under the information function with bounds given by $\left(\hat{\theta}_l, \hat{\theta}_u\right)$.

For Fisher information with a posterior weight function (FIP), the weight function in Equation 19 is the posterior density of the provisional ability estimate $\hat{\theta}$, and the integration is performed over the entire range of the ability continuum. Thus, for FIP, this integration becomes

$$FIP_j\left(\theta\right) = \int_{-\infty}^{\infty} p\left(\theta \middle| u_1, u_2, \ldots, u_{j-1}\right) I_j\left(\theta\right) d\theta \qquad \text{(Eq. 22)}$$

where $p\left(\theta \middle| u_1, u_2, \ldots, u_{j-1}\right)$ is the posterior density for the estimate $\hat{\theta}$ after administering $j-1$ items.

Methods based on Fisher information at a point estimate $\hat{\theta}$ or an interval about that estimate may be classified as local information methods. There is an implicit assumption in these procedures that the provisional estimate $\hat{\theta}$ is in the neighborhood of an examinees true ability θ_0. Chang & Ying (1996) argue that early in a CAT administration, when the number of items administered is small, "the estimator [$\hat{\theta}$] may not be close to θ_0, in which case the information inside a small region around [$\hat{\theta}$] would not be useful." It turns out that this argument is not new; Lord & Novick (1968, sect. 16.5) described it in terms of an "attenuation paradox", as pointed out by van der Linden & Pashley (2000). The problem is that item selection procedures that disregard the possibility that $\hat{\theta}$ is not in the neighborhood of θ_0 will choose optimal items with respect to the estimate $\hat{\theta}$ but not with respect to θ_0, the quantity of most interest.

Chang & Ying (1996) propose that a global information measure, rather than a local one, be used for item selection, especially at the early stages of a CAT. This global information measure, or Kullback-Leibler (KL) information, does not impose the restriction that $\hat{\theta}$ be close to θ_0, as in local information methods. Kullback-Leibler (KL) information is obtained through the Neyman-Pearson likelihood ratio method. Chang & Ying (1996) note that this method is optimal for testing $\theta = \theta_0$ versus $\theta = \theta_1$, where θ_0 and θ_1 are two points on the ability continuum. Thus, while local information methods are concerned with the information content of an item at one ability point (and possibly the error associated with that point, as in the case of FII and FIP), global information methods are concerned with the information content of an item with respect to two ability points. Chang & Ying (1996, p. 217) explain that KL information "is a function of two levels, [θ_0 and θ_1]" and that it "represents the discrimination power of the item on the two levels." That is, KL information is a multivariate function, whereas local information (i.e., Fisher information) is a univariate function.

If θ_0 is an examinee's true ability, and $\hat{\theta}$ is the provisional estimate, then the KL item information for item *j* is

$$K_j\left(\hat{\theta},\theta_0\right) = E\left[\log \frac{L\left(\theta_0 \middle| U_j\right)}{L\left(\hat{\theta} \middle| U_j\right)}\right] \qquad \text{(Eq. 23)}$$

where the expectation is over the response variable U_j. Since the likelihood for a (dichotomous) item *j* is given by

$$L\left(U_j \middle| \theta\right) = P_j^{U_j}(\theta)\left[1 - P_j(\theta)\right]^{1-U_j} \qquad \text{(Eq. 24}$$

the KL item information is then

$$K_j\left(\hat{\theta},\theta_0\right) = P_j\left(\theta_0\right)\log\frac{P_j\left(\theta_0\right)}{P_j\left(\hat{\theta}\right)} + \left[1 - P_j\left(\theta_0\right)\right]\log\frac{P_j\left(\theta_0\right)}{P_j\left(\hat{\theta}\right)} \qquad \text{(Eq. 25)}$$

The properties of the KL information follow. First, it is not symmetric, so $K_j\left(\hat{\theta},\theta_0\right) \neq K_j\left(\theta_0,\hat{\theta}\right)$. Second, $K_j\left(\hat{\theta},\theta_0\right) \geq 0$. Third, if $\hat{\theta} = \theta_0$, then $K_j\left(\hat{\theta},\theta_0\right) = 0$. KL information then describes the power of an item *j* to discriminate between two points on the ability scale, $\hat{\theta}$ and θ_0. If in fact these points are equal, as in the third property, the item cannot discriminate between two points on the ability scale, since only one point is represented.

The immediate problem with KL information in a practical testing situation is that the true ability θ_0 is unknown. Chang & Ying (1996) propose to integrate this unknown parameter out to calculate an index. Their procedure builds on the GWIC shown in Equation 19. The resulting integral is

$$KL_{index,j} = \int_{\hat{\theta}-\delta_n}^{\hat{\theta}+\delta_n} K_j\left(\hat{\theta},\theta\right)d\theta \qquad \text{(Eq. 26)}$$

where $\hat{\theta}$ is the provisional ability estimate, $K_j\left(\hat{\theta},\theta\right)$ is the KL information, and $\delta_n = \frac{z}{\sqrt{n}}$, a decreasing function in *n*, the number of items administered, and *z* is a normal deviate selected at the desired level of confidence. Chang & Ying (1996) also propose a Bayesian information index, similar to the FIP in mathematical form (see Equation 22). This KL information method with a posterior weight function (abbreviated KLP) is given by

$$KLP_{index,j} = \int_{-\infty}^{\infty} p\left(\theta \middle| u_1, u_2, \ldots, u_{j-1}\right) K_j\left(\hat{\theta},\theta\right)d\theta \qquad \text{(Eq. 27)}$$

where $p\left(\theta \middle| u_1, u_2, \ldots, u_{j-1}\right)$ is the posterior density for the estimate $\hat{\theta}$ after administering $j-1$ items.

Because an examinee's true ability is likely not to be well estimated at the early stages of a CAT (i.e., when the number of administered items is small), Chang & Ying (1996) suggest that global information be used. However, once many items have been administered and the estimated ability $\hat{\theta}$ converges to the true ability θ_0, they suggest that local information be used. It should be noted that the global information item selection procedures are the most computationally burdensome of the methods described so far. Thus, these procedures may have limited applicability to practical testing situations where items must be selected and administered to examinees in an expedient manner. How well these methods perform as compared to the local information methods will be explored in the fifth section, evaluating item selection procedures.

Operational constraints in item selection

An operational CAT places constraints on item selection beyond psychometric criteria such as item information. Although a CAT should select items that are targeted at an examinee's ability, an operational test (adaptive or otherwise) must also conform to test blueprint specifications. These constraints must then be handled by the item selection procedure. Now the problem of item selection has become more complicated, as the next "optimal" item to choose for examinees is not only a function of item information, but also of blueprint specifications.

Adding further complexity to item selection in an operational CAT is exposure control. Unlike tests which are administered periodically (e.g., once every four months), a CAT may be administered on a continuous basis, weekly or even daily. Whereas item pools used for periodic tests may be partially or completely refreshed in time for the next administration, it is practically infeasible to refresh item pools for CATs in this manner. Thus, the added convenience of continuous testing afforded by CATs is also its greatest drawback, in that items may become compromised on time scales that are rather short.

The operational CAT item selection procedure is constrained by at least three requirements: (1) select the item which is most appropriate for an examinee's ability; (2) insure that the entire set of items administered to an examinee is balanced according to test specifications; and (3) insure that the items administered to an examinee have not been exposed to other examinees too frequently. Popular techniques for satisfying these three requirements typically divide the problem into two parts: satisfy (1) and (2) together, then satisfy (3). Techniques for satisfying item information selection and test specification balancing include Kingsbury & Zara's (1991) constrained computerized adaptive test (CCAT), and Stocking & Swanson's (1993) and Swanson & Stocking's (1993) weighted deviations model (WDM). Techniques for controlling item exposure include those of McBride & Martin (1983) and Sympson & Hetter (1985).

To insure that test specifications are met in a CAT, Kingsbury & Zara's (1991) CCAT employs a strategy of monitoring for each examinee the proportion of items selected from various content areas. Content balancing is fixed before CAT administration by assigning targeted content area proportions (e.g., a science test might require 50% of items from the biological sciences, and 50% from the physical sciences). Before the next item is selected for an examinee, an accounting of the items already administered to that examinee is performed and the observed proportions with respect to content area are calculated. If the proportion for a content area is below its targeted proportion, selection of the next item is restricted to the items within that content area. Maximum item information is used to select the next item from the restricted set.

Stocking & Swanson (1993) and Swanson & Stocking (1993) developed the weighted deviations model (WDM) to handle practical test assembly constraints including but not limited to content balancing. As described by Eignor, Stocking, Way, & Steffen (1993), the philosophy underlying their approach is different from approaches such as Kingsbury & Zara's. Regarding WDM, they note, "Thus constraints, including statistical constraints, are thought of as more like 'desired properties' than as true constraints. This approach recognizes the possibility of constructing a test that may lack all of the desired properties at the expected levels, but emphasizes the minimization of aggregate failures" (p. 5). Further, Stocking & Lewis (1998) comment that "in the WDM, the item pool is ordered by employing a methodology from the decision sciences that models the behavior of expert test specialists" (p. 59). Thus, the weights in the WDM refer to the relative importance of test form (or for a CAT, tailored test administration) attributes as determined by test specialists (Eignor et al., 1993). The goal of the WDM is then to minimize the weighted deviations from the nominal levels.

Controlling item exposure is more a security concern than a psychometric one. Nevertheless, it is important because, as Stocking & Lewis (2000) note, item exposure increases the risk that examinees will obtain "preknowledge" about tests and their constituent items through the sharing of information (p. 164). Though advantageous from a security standpoint, controlling item exposure comes at the expense of efficiency, as it overrides an item selection procedure's best choice for the next item to administer to an examinee (Thissen & Mislevy, 1990). As Eignor et al. (1993) note, item exposure control forces CAT administrations to be longer, but these "longer tests may be viewed as a reasonable exchange for greater item and test security" (p. 8).

McBride & Martin (1983) proposed a randomization technique to control item exposure. The logic behind their technique is that early in the CAT, examinees are not well differentiated and are likely to receive nearly the same set of items as selected by maximum information selection. Rather than administering the single best item at these early stages, a group of appropriate items is created and an item from this group is selected at random. For example, this group might include the best item, as measured by maximum item information, as well as a few of the next-best items. The test begins with an item selected at random from a group of items of size five; the next item is selected from a group of size four, and the process continues until the fifth item, where the group is of size one and hence the item with maximum information is selected. From the fifth item onward, the randomization technique is no longer employed (Thissen & Mislevy, 1990). The primary advantage of McBride & Martin's technique is that it is rather easy to employ. The disadvantage, however, is that item exposure is controlled indirectly; that is, the probability that any given item will be administered is not known in advance (Stocking & Lewis, 1995).

To control for item exposure more directly, Sympson & Hetter's (1985) technique is grounded in a probabilistic framework. Attached to each item is the probability $P(A,S)$, where S is the event that the item is selected, and A is the event that the item is administered. Thus, the probability $P(A,S)$ that an item is selected and administered is given by

$$P(A,S) = P(A|S)P(S) \qquad \text{(Eq. 28)}$$

where $P(S)$ is the probability that an item is selected (e.g., by maximum information item selection), and $P(A|S)$ is the probability that an item is administered, given that it is selected.

To utilize the Sympson & Hetter technique, a desired level of item exposure, say *r*, where $r \in [0,1]$ is chosen for all items. Then for all items, it is desired that $P(A,S) \le r$. An item exposure control parameter *k* is chosen for each item such that $k = P(A|S)$. Values for *k* are found through iterative simulations (Stocking & Lewis, 1995, p. 10). Once the parameters *k* have been identified, item exposure control proceeds as follows: (1) select the next most appropriate item for administration and obtain its exposure control parameter *k*; (2) generate a random deviate *u* from *UNIF*(0,1); (3) if $u \le k$, administer the item, otherwise, do not administer the item, remove it from the pool of potential items for the examinee, and go back to step 1 (Eignor et al., 1993).

The advantage of the Sympson & Hetter technique is that the probabilities of item exposure are controlled directly; however, it may not be possible to obtain stable exposure control parameters *k* for each item through the iterative simulations (Stocking & Lewis, 1995). Further, these parameters are dependent on the item pool and the manner in which the test administration simulations are conducted, and hence new simulations must be conducted when changes are made to the item pool or administration procedures (Stocking & Lewis, 1998).

In addition to content balancing and item exposure control, operational item selection procedures must also take into account the possibility that locally-dependent items may need to be administered. Most commonly, these are items that are grouped into contextually-related item sets; for example, a passage-based item followed by a number of items relating to that passage. One possible solution to overcoming the dependency among these items is to group these items together into a single unit, then model the responses to these items as multiple levels of a polytomous IRT model. These units, sometimes called testlets, would then be administered to an examinee and scored according to an appropriate polytomous IRT model (Wainer & Mislevy, 1990). For example, responses to a testlet of four items could be collapsed into one of five score categories, zero through four, according to the number correct score for all items within the testlet. Then a polytomous IRT model, such as Samejima's (1969) graded response model or Master's (1982) partial-credit model, may be used to model item responses. CAT item selection procedures for these testlets would mirror those for dichotomous items, since information functions may be computed for polytomous items as well as dichotomous ones.

Evaluation of item selection procedures

Item selection in an operational CAT is a multifaceted process, whereby selection must take into account statistical criteria (such as item information), test specifications, and exposure control. Most studies on item selection procedures have focused on one aspect of this process, holding the others constant or disregarding them altogether. Studies by Chang & Ying (1996), Chen, Ankenmann, & Chang (2000), and Cheng & Liou (2000) explore the issue of global versus local information through simulation studies, but do not include content balancing or exposure control as variables in their study. Kingsbury & Zara (1991) investigate a procedure for content balancing under maximum information item selection, but also do not employ exposure control in their simulations. Operational concerns are the focus of Eignor, Stocking, Way, & Steffen's

(1993) study, where content balancing and exposure controls are investigated, but the maximum information criterion is used for all conditions of the study.

Chang & Ying (1996) proposed a method of global information item selection using the Kullback-Leibler (KL) information, and compared it with local information, or Fisher information (FI), item selection procedures. They conducted two simulation studies, each with different item pools and different test lengths. In the first of their studies, item parameters *a*, *b*, and *c* for a 3P model were generated from uniform distributions, such that *a* ~ *UNIF*(0.5, 2.5), *b* ~ *UNIF*(-3.6, 3.6), and *c* ~ *UNIF*(0.0, 0.25). The size of the pool was 800 items, and a fixed-length CAT of 14 items was administered. In their second study, the item parameters were obtained from a 1992 NAEP Reading Assessment, with a pool of 254 items. Of these items, 122 were modeled with a 2P model, and the remainder with a 3P model. Again, a fixed-length CAT was administered, but with a length of 40 items. Each study simulated responses for 1000 examinees. Both bias and mean-squared-errors (MSE) were examined for the two studies.

For their first simulation study, Chang & Ying (1996) found that the KL information method resulted in less bias and smaller MSE than the FI method for examinees at the lower extremes of ability, $\theta = \{-3,-2,-1.5\}$. Both KL and FI performed similarly at ability levels $\theta > -1.5$. For their second simulation study, they again found less bias at the lower extremes of ability, $\theta = \{-2,-1\}$, but smaller MSE only for $\theta = -2$. When the number of items administered was small, however, KL resulted in smaller MSE than FI.

Fan & Hsu (1996) also explored global information item selection procedures, but their item pool was significantly smaller than that used by Chang & Ying (1996). Whereas Chang & Ying (1996) used an item pool of size 800 for their first study, and 254 for their second study, Fan & Hsu (1996) limited their item pool size to 100 items. The results of Fan & Hsu's (1996) study were not consistent with those observed by Chang & Ying (1996); that is, there was no difference between KL and FI item selection in terms of bias and MSE of ability estimates.

Chen, Ankenmann, & Chang (2000) extended Chang & Ying's (1996) study to investigate additional local and global information-based item selection procedures. Whereas Chang & Ying (1996) limited the local information selection method to maximum Fisher information, and the global information selection method to the maximum KL information index (as calculated by Equation 26), Chen et al. (2000) examined Fisher interval information (FII, see Equation 19), Fisher information with a posterior weight function (FIP, see Equation 22), and Kullback-Leibler information with a posterior weight function (KLP, see Equation 27). Reflecting on the findings of Fan & Hsu's (1996) study, Chen et al. (2000) posit that the benefits of KL information may be reduced in cases of smaller item pools; further, they suggest that the item characteristics themselves may play a role in the utility of KL over FI.

Chen et al. (2000) conducted two simulation studies to investigate the FI, FII, FIP, KL, and KLP item selection procedures. In their first study, item parameters a, b, and c for a 3P model were generated as $a \sim N(1, 0.25)$, $b \sim U(-3.6, 3.6)$, and $c \sim U(0.0, 0.3)$; the size of this item pool was 400. The item pool for their second study was identical to the item pool used for Chang & Ying's (1996) second study (i.e., the 1992 NAEP Reading

Assessment items). Both studies simulated responses for 1000 examinees administered a 20-item CAT. Their study examined bias, standard errors, and root-mean-squared errors (RMSE).

Chen et al. (2000) found that of three research hypotheses, only two were supported by their study. These two hypotheses were: (1) that early in a CAT administration, FI would perform best at ability values near $\theta = 0$; and (2) that as the number of items administered in the CAT increased, all item selection procedures would perform similarly. Their hypothesis that FII, FIP, KL, and KLP would perform better than FI at the early stages of testing for ability values far from $\theta = 0$, was supported only at the lower extremes of ability, where $\theta = \{-3,-2\}$. They write, "Differences among [these methods] with respect to bias, RMSE, and SE...were negligible for tests of more than 10 items….For tests longer than 10 items, there appeared to be no precision advantage of one [method] over another" (p. 253).

Cheng & Liou (2000) also investigated KL and FI item selection procedures, and like Chang & Ying (1996) and Chen et al. (2000), their simulation study used items from a 1992 NAEP Reading Assessment. However, rather than using all 254 items, they limited the size of their item pool to 204. Their study simulated responses for 1000 examinees administered a 30-item CAT. Dependent variables in their study included bias and MSE. Their findings paralleled those of Chang & Ying (1996) and Chen et al. (2000), in that the KL and FI item selection procedures performed similarly for more than 10 items. Cheng & Liou (2000) also include some measures of processing time; they note that the more computationally-intensive KL procedure was on average 60 times slower than the FI procedure.

Content balancing in the absence of exposure control was investigated by Kingsbury & Zara (1991), in order to examine their constrained CAT (CCAT) item selection procedure. Their simulation study included: (a) two examinee groups of size 10,000, where one group was distributed as $N(0, 1)$ and the other as $N(1.5, 1)$; (b) three item pools of sizes 100, 300, and 500; and (c) two CAT administrations, unconstrained (i.e., not content-balanced) and constrained (CCAT). Each CAT administration was fixed to a length of 48 items, and items were assigned to four content areas as follows: one-third from content area A, one-third from content area B, one-sixth from content area C, and one-sixth from content area D. For a desired precision of measurement, they found that the CCAT required between 5% and 11% more items than an unconstrained CAT. They note that the cost of content balancing comes at the expense of test length; i.e., a longer content-balanced CAT is required to achieve the same level of precision as the corresponding unconstrained CAT.

Satisfying operational constraints was the primary focus of the study by Eignor, Stocking, Way, & Steffen (1993). They examined CAT versions of five different tests: the SAT Verbal, SAT Mathematics, GRE Verbal, GRE Quantitative, and GRE Analytical. Although they utilized simulation techniques to conduct their study, all work was performed in an operational context; that is, the study was part of Educational Testing Service's effort to make available CAT versions of their paper-and-pencil tests. Their results included not only an evaluation of psychometric, content balancing, and item exposure criteria, but also a review by test specialists.

Eignor et al. (1993) utilize the weighted deviations model (WDM) for content balancing in conjunction with FI item selection. For exposure control, they employ a count-down randomization technique starting with eight items for the SAT simulations, and the extended Sympson & Hetter technique for the GRE simulations. (Note that Stocking (1993) developed the extended Sympson & Hetter technique to work with item stimulus material, such as passages, as well as the items that follow.) Responses for between 100 and 200 examinees were generated per ability scale point, which varied between 9 for the GRE Analytical and 19 for the SAT Verbal.

All of their simulated CATs used a fixed-length stopping rule, though the length of the CAT depended upon the particular test. Further, Eignor et al. (1993) found that operational constraints could be satisfied when item pools were approximately 12 times larger than the length of the CAT. For example, the SAT Verbal simulations were fixed at 27 items and the SAT Mathematics at 20 items; item pools of size 303 and 235, respectively, satisfied the constraints. For a 30-item GRE Verbal, a 28-item GRE Quantitative, and a 35-item GRE Analytical, item pools of size 350, 330, and 449, respectively, satisfied the constraints.

Eignor et al. (1993) found that all five of the CATs met or exceeded targeted values for test form reliability. Further, they found that content balancing was for the most part satisfied, with some violations for those aspects receiving smaller weights in the WDM. With respect to item selection, it is worth noting that none of the CATs used all items in the pool. On average, 84% of the items in the pool were administered. Item exposure rates varied depending on the control methodology. For the SAT simulations, where the count-down randomization technique was used, the highest exposure rates were between 0.5 and 0.6. The average exposure rate was approximately 0.11. For the GRE simulations, where the extended Sympson & Hetter technique was used, the highest exposure rates were between 0.2 and 0.3. The average exposure rate was approximately 10%.

Ultimately, CATs should conform to what expert test specialists would require were they to assemble the tests themselves. Eignor et al. (1993) generated paper-and-pencil copies of CAT administrations generated by their simulations and submitted them to test specialists for review. These specialists were not made aware of the constraints utilized in the simulations, nor were they told the ability levels for which the individual tests were targeted. After review, they did identify some problems with the CATs, but these problems arose not from the item selection procedures, but rather from characteristics of the item pools. For example, they identified problems with those forms designed for examinees at the extremes of ability. Eignor et al. (1993) attribute this problem to a lack of items in the pool tailored to such examinees.

CHAPTER 3

Methodology

The following discussion of the methodology employed in this study is organized into four major sections: (1) the efficiency of item selection in the context of computerized adaptive testing (CAT); (2) an alternative ability estimation procedure based on hypothesis-testing; (3) an overview of the experimental design; and (4) the CAT simulation methods used to carry out the study. Within the second section, a brief review of related methods, particularly hypothesis-tests based on person-fit statistics, is included in addition to the mathematical development of the alternative item selection procedure.

Defining the efficiency of item selection in the context of CAT

The evaluation criterion upon which this study relies is a precisely-defined measure of efficiency. An argument for this definition was provided in the introductory chapter; the same argument is presented here along with additional points related to the methodology of the study. The precise definition of efficiency in this context begins with the statistical concept of relative efficiency, which is then used to define an efficiency measure for the items selected by a CAT.

To begin, consider two tests, A and B, administered to an examinee possessing a true ability θ. The precision with which this examinee may be measured by test A is given by the accumulated test information, as defined in Equation 12 and shown to be equivalent to the sum of the information provided by each item administered, as given by Equation 18. The expression $I_A^{(T)}(\theta)$ may then indicate the accumulated test information for test A at the examinee's true ability θ, and similarly $I_B^{(T)}(\theta)$ for test B. The relative efficiency of test A over test B is then

$$RE(A,B|\theta)=\frac{I_A^{(T)}(\theta)}{I_B^{(T)}(\theta)} \tag{Eq. 29}$$

If test A is more efficient than test B at true ability θ (i.e., test A yields more precise measurements), then $RE(A,B|\theta)>1$. Conversely, if test B is more efficient at θ, $RE(A,B|\theta)<1$. In the case where both tests are equally efficient at θ, $RE(A,B|\theta)$ is exactly equal to one.

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

$$RE\left(CAT, optimal \middle| \theta\right) = \frac{I_{CAT}^{(T)}(\theta)}{I_0^{(T)}(\theta)} \qquad \text{(Eq. 30)}$$

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.

Spanos (1999, p. 609) maintains that since relative efficiency alone is not necessarily useful (e.g., a poor estimator is relatively more efficient than an even poorer estimator), some fixed point of reference, namely the Cramer-Rao lower bound for the variance of an estimator, is required. In IRT applications where ability estimates are obtained by maximum likelihood (ML), the asymptotic variance of the ML estimator is in fact equal to the Cramer-Rao lower bound, or $1/I^{(T)}(\theta)$, where $I^{(T)}(\theta)$ is the accumulated test information at the true ability θ. However, the specific items contributing to $I^{(T)}(\theta)$ for an individual examinee being administered a CAT are not fixed, and hence an additional lower bound—this one resulting from the choice of items—is needed to define efficiency. It is the quantity $I_0^{(T)}(\theta)$ described above that sets the lower bound on the variance for an estimator of θ (or, an upper bound on the precision of such an estimator) in the CAT environment with an item pool of finite size. Thus, the ratio $I_{CAT}^{(T)}(\theta)/I_0^{(T)}(\theta)$ is not simply a measure of relative efficiency, but of efficiency itself, and may be used to define the efficiency of a CAT. The primary advantage of this definition is that item selection procedures may all be compared to a fixed point of reference, one that characterizes the most efficient estimator possible.

Using the present definition of $I_0^{(T)}(\theta)$ in an evaluation of item selection procedures may be somewhat unrealistic, for the simple reason that a CAT must select an arbitrary starting point for the first item draw. Quite often, the mean of an ability distribution, either assumed a priori or in fact observed from previous test administrations, is taken as the initial ability estimate for examinees. If such a method is utilized, then any subset of items administered to an examinee must necessarily contain an item drawn based on this starting estimate. Such arbitrariness could conceivably penalize the efficiency measure, as calculated by Equation 30, for an item selection procedure.

In order to produce a more appropriate measure, the definition of $I_0^{(T)}(\theta)$ may be modified such that this first item is always included in the "optimal" subset, recognizing that it may or may not be among the most informative items in the pool for an examinee with true ability θ. For this study, the following (modified) definition of $I_0^{(T)}(\theta)$ is used.

Suppose that *J* items are administered to an examinee, where *j* = 1, 2, …, *J*, and $I_j(\theta)$ is the item information for item *j* at ability θ. Then $I_0^{(T)}(\theta)$ is defined as

$$I_0^{(T)}(\theta) = I_1(\theta) + \sum_{j=2}^{J} I_j(\theta) \qquad \text{(Eq. 31)}$$

where $I_1(\theta)$ is the information at θ provided by the first item, and $\sum_{j=2}^{J} I_j(\theta)$, which is the sum of the information from the remaining $J-1$ items, is not exceeded by any other combination of $J-1$ items drawn from the item pool. That is, with $I_1(\theta)$ fixed at the start of the test, no other subset of $J-1$ items yields a greater measure of accumulated test information at θ. This modification recognizes that one degree of freedom is lost in the item selection process because the first item is arbitrarily drawn. Hence, efficiency in this study is defined as in Equation 30, but with a modified upper bound on the precision of measurement $I_0^{(T)}$ given by Equation 31, such that

$$Efficiency\left(CAT|\theta\right) = \frac{I_{CAT}^{(T)}}{I_0^{(T)}} = \frac{I_{CAT}^{(T)}}{I_1(\theta) + \sum_{j=2}^{J} I_j(\theta)} \qquad \text{(Eq. 32)}$$

An alternative ability estimation procedure based on hypothesis-testing

Item selection in CAT is influenced by two procedures: item selection and ability estimation. With respect to item selection, maximum Fisher information (FI) item selection is the most common; for ability estimation, ML, MAP, and EAP are most common. The majority of research on improving the efficiency of item selection has concentrated on item selection procedures, and a number of alternative item selection procedures have been proposed. Four of these procedures—Fisher interval information (FII), Fisher information with a posterior weight function (FIP), Kullback-Leibler (KL) information, and KL information weighted by a posterior density (KLP)—have been proposed as reviewed in Chapter 2. However, the efficiency of item selection might also be improved by considering alternative ability estimation procedures. The golden section search (GSS) strategy, also reviewed in Chapter 2, is one example of an alternative ability estimation procedure.

Proposed here is a new alternative ability estimation procedure, fundamentally different from those previously reviewed. The procedure utilizes a hypothesis-testing approach in conjunction with a conventional ability estimation procedure such as ML or MAP, whereby a hypothesis test is constructed and a decision rule is followed in order to select the next item for administration. The procedure is developed according to the following, which are then discussed in turn: (1) statement and interpretation of the null and alternative hypotheses; (2) derivation of the test statistic; (3) decision rule and subsequent item selection. In addition, because some aspects of this procedure are similar to those used in person-fit statistics, a comparison of this procedure with these methods is provided.

Statement and interpretation of the null and alternative hypotheses

The null hypothesis in this procedure is similar to those used for tests of model fit. In these tests, a model is proposed and the observed data is compared to what is expected under the model. If the observed data differ significantly from the model expectations, then the null hypothesis is rejected. In such a case, the conclusion is typically that the model does not fit the data. On the other hand, if sufficient evidence does not exist for rejecting the null hypothesis, the model is said to fit the data[5].

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

[5] It may be argued that when the null hypothesis is not rejected, this interpretation is improper. Nevertheless, it is often the only type of evidence available for making decisions about model-data fit.

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 a dichotomously-scored 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, where information is given by Equation 17. 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,

$$\theta_{\max,1} = \theta_{\max,2} = \cdots = \theta_{\max,N} = \theta \qquad \text{(Eq. 33)}$$

The next relationship links Equation 33 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 33. This expected proportion correct, or p, is then defined as

$$p = \frac{\sum_{i=1}^{N} P(U_i = 1|\theta_{\max,i})}{N} \qquad \text{(Eq. 34)}$$

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

$$\hat{p} = \frac{\sum_{i=1}^{N} X_i}{N}, \quad X_i = \{0,1\} \qquad \text{(Eq. 35)}$$

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 to be fixed at some true value θ, this decision further suggests that the relationship in Equation 33 be retained[6]. 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 33; 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.

Derivation of the test statistic

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

[6] If items are perfectly targeted at examinee ability, then Equation 34 follows by deduction. However, the inductive step is somewhat more involved. Satisfying Equation 34 is a necessary but not sufficient condition for concluding Equation 33. Caution must be exercised in interpreting model-data fit under retention of the null hypothesis. Nevertheless, if Equation 34 is not satisfied (i.e., when the null is rejected), it cannot be the case that Equation 33 is true.

identically distributed. (Thus, the assumption of local independence is assumed here.) The proportion correct for X_i (or, the mean of the X_i) may then be defined as

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad \text{(Eq. 36)}$$

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

$$p = E[\hat{p}] = E\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{1}{n} E\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{\sum_{i=1}^{n} p_i}{n} \qquad \text{(Eq. 37)}$$

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

$$Var[\hat{p}] = Var\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{1}{n^2} Var\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var[X_i] = \frac{\sum_{i=1}^{n} p_i q_i}{n^2} \qquad \text{(Eq. 38)}$$

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

$$z^* = \frac{\hat{p} - p}{\sqrt{Var[\hat{p}]}} \qquad \text{(Eq 39)}$$

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 37 and 38 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 33. Thus, under the 3P model, the p_i for an item *i* used in these equations are given by

$$p_i = P\left(X_i = 1 \middle| \theta_{\max,i}\right) = c_i + \frac{(1 - c_i)}{\left[1 + \exp\left(-Da_i\left(\theta_{\max,i} - b_i\right)\right)\right]} \tag{Eq 40}$$

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

$$\theta_{\max,i} = b_i + \frac{1}{Da_i}\ln\left(\tfrac{1}{2} + \tfrac{1}{2}\sqrt{1 + 8c_i}\right) \tag{Eq 41}$$

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 41 into Equation 40 results in the following simplification for p_i

$$p_i = P\left(X_i = 1 \middle| \theta_{\max,i}\right) = c_i + (1 - c_i)\left[1 + \frac{2}{1 + \sqrt{1 + 8c_i}}\right]^{-1} \tag{Eq 42}$$

Using this expression for p_i, the necessary quantities $E[\hat{p}]$ and $Var[\hat{p}]$ may be calculated by means of Equations 37 and 38.

Note that under 1P and 2P IRT models, if $c_i = 0$, then $p_i = 0.5$. Under the 3P model, when $c_i > 0$, $p_i > 0.5$. As discussed earlier, this is the expected proportion when a CAT is targeting items at an examinee’s true ability level.

Distribution of the test statistic under the null hypothesis

Under the null hypothesis, it is assumed that all items administered to an examinee possess maximum information at that examinee’s true ability θ, as given in

Equation 33. The hypothesis-testing procedure assumes that the distribution of the test statistic z^* under the null hypothesis follows a $N(0,1)$ distribution. However, the assumption that $z^* \sim N(0,1)$ is an asymptotic result, and may not hold for relatively short tests (i.e., tests less than 30 items in length).

In order to determine to what extent z^* follows a $N(0,1)$ distribution under the null hypothesis, a simulation was conducted. Three levels of ability were chosen such that $\theta = \{-2, 0, +2\}$, with 1000 replications for each ability. For all items i administered to examinees, $\theta_{\max,i} = \theta$, as in Equation 33. The empirical sampling distribution for z^* was examined for tests of length 5, 10, 15 and 25. For each combination of examinee true ability θ with test length, the mean and standard deviation of the distribution is reported in Table 1. In addition, the type I error rates corresponding to the nominal α-levels 0.05, 0.10, and 0.20 for a two-tailed test under the $N(0,1)$ distribution are reported in Tables 2 through 4. In the case of the $N(0,1)$ distribution, the critical z-value for $\alpha = 0.05$ is $z_c = 1.960$; for $\alpha = 0.10$, $z_c = 1.645$; and for $\alpha = 0.20$, $z_c = 1.282$. Type I error rates were then computed as the frequency with which $|z^*| > z_c$ for each condition.

Table 1. Mean and standard deviation of z^* under the null hypothesis.

	Mean				Standard deviation			
True ability θ	5 items	10 items	15 items	25 items	5 items	10 items	15 items	25 items
−2	0.020	0.010	0.020	0.013	0.99	1.01	0.98	1.00
0	0.01	0.0045	0.0015	-0.029	1.02	1.03	1.02	0.99
+2	-0.0083	0.0045	0.0015	0.029	1.00	1.00	1.01	1.02

Table 2. Type I error rates for z^*, with nominal $\alpha = 0.05$ ($z_c = 1.960$).

	Test length			
True ability θ	5 items	10 items	15 items	25 items
−2	0.006	0.049	0.055	0.027
0	0.009	0.059	0.066	0.034
+2	0.009	0.042	0.057	0.041

Table 3. Type I error rates for z^*, with nominal $\alpha = 0.10$ ($z_c = 1.645$).

	Test length			
True ability θ	5 items	10 items	15 items	25 items
−2	0.158	0.101	0.103	0.095
0	0.176	0.115	0.121	0.098
+2	0.162	0.093	0.103	0.108

Table 4. Type I error rates for z^*, with nominal $\alpha = 0.20$ ($z_c = 1.282$).

	Test length			
True ability θ	5 items	10 items	15 items	25 items
−2	0.158	0.181	0.179	0.231
0	0.176	0.206	0.207	0.214
+2	0.162	0.197	0.199	0.213

From Table 1, the means and standard deviations of the empirical sampling distributions for z^* are consistent with a distribution with a mean of 0 and a standard deviation of 1. The type I error rates shown in Tables 2 through 4 are for the most part consistent with the nominal type I error rates for a $N(0,1)$ distribution, with the only major departures occurring for tests of 5 items in length. In the case of the nominal $\alpha = 0.05$ and a test length of 5 items, the observed type I error rates are substantially less than expected, whereas for the nominal $\alpha = 0.10$, the type I error rates are substantially greater than expected. Note that the observed type I error rates for the nominal $\alpha = 0.20$ are identical to those observed for $\alpha = 0.10$. This equality of type I error rates for different values of z_c is due to the discreteness of the distribution of z^* for very short test lengths. (For a test of 5 items in length, only 6 values of z^* are possible.) Although the convergence of z^* to $N(0,1)$ is an asymptotic result, it appears that this normal approximation for z^* is accurate for tests of 10 items in length or greater.

Decision rule and subsequent item selection

Equation 39 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., $\left|z^*\right| > z_c$). 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. 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 $E[p]$. Thus, a decision to reject the null hypothesis when $\hat{p} < E[p]$ is equivalent to stating that $\hat{p}$ lies outside the confidence interval for $E[p]$, and specifically, beyond its lower confidence limit of $E[p] - z_c\sqrt{Var[p]}$. Likewise, rejection of the null when $\hat{p} > E[p]$ demands that $\hat{p}$ must lie beyond the upper confidence limit $E[p] + z_c\sqrt{Var[p]}$.

It is then reasonable to suppose in this situation that, for $\hat{p} < E[p]$, the location of the alternative distribution $E[\hat{p}]$ is less than or equal to $E[p] - z_c\sqrt{Var[p]}$, and for $\hat{p} > E[p]$, the location of the distribution of $E[\hat{p}]$ is greater than or equal to $E[p] + z_c\sqrt{Var[p]}$. Then an approximation to $E[\hat{p}]$ may be denoted by $\hat{p}^*$, such that

$$\begin{aligned} \hat{p}^* &= E[p] + z_c\sqrt{Var[p]}, \quad \hat{p} > E[p] \\ \hat{p}^* &= E[p] - z_c\sqrt{Var[p]}, \quad \hat{p} < E[p] \end{aligned} \qquad \text{(Eq. 43)}$$

where each of the quantities $E[p]$, $Var[p]$, 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 ICC for a group of items is a monotonically increasing function of θ. Thus, any proportion $P(X = 1|\theta)$ corresponds to one and only one θ. However, to insure an inverse transformation of proportions to ability values, it is also necessary to insure that every element in the range of the proportions can be mapped to elements in the domain of the ability values. For the 3P IRT model, it is not true in general that for any $p \in [0,1]$, there exists a $\theta \in (-\infty, \infty)$ such that $p = P(X = 1|\theta)$. For example, if a guessing parameter $c > 0$ is present, then any $p < c$ will not be mapped into the domain of θ. Thus, a uniquely specified inverse transformation of proportions p to ability values θ does not exist under all circumstances. However, in those cases where the inverse transformation fails, remedial measures may be taken. The specific procedures for transforming $\hat{p}^*$ to $\hat{\theta}^*$ are considered next.

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, since the TCC is the sum of the ICCs for each administered item. 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

$$\overline{P}(X=1|\theta) = \frac{\sum_{i=1}^{N} P(X_i = 1|\theta)}{N} \qquad \text{(Eq. 44)}$$

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 $\hat{p}^* \le \overline{P}(X=1|\theta=c)$, 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.

One advantage of this method of halving is that a solution is always guaranteed, even under anomalous circumstances. For example, if the obtained $\hat{p}^*$ is less than the average ICC function for the lower bound, that is $\hat{p}^* < \overline{P}(X = 1|\theta = -4)$, the procedure will return a limiting solution of $\theta = -4$. These limiting solutions are sufficient for the CAT algorithms, since finite bounds are placed on the ability continuum in any event. Once the new provisional ability estimate $\hat{\theta}^*$ is obtained, an item selection procedure uses this new provisional ability estimate to select the next item for administration.

Identifying optimal z_c values

In order to use the decision rule discussed in the previous section, the alternative ability estimation procedure must employ 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. That is, when the procedure is invoked, it should meet or exceed the outcomes (e.g., test efficiency) obtained by using the conventional ability estimation procedure, and should not perform less successfully than the conventional procedure.

This value z_c was determined empirically as described in Appendix B, with a summary of the empirical procedure discussed here. Two optimal z_c values were identified; one for ML estimation concurrent with the alternative ability estimation procedure (hereunto denoted as ML/Alt), and the other for MAP estimation concurrent with the alternative ability estimation procedure (hereunto denoted as MAP/Alt).

Optimal z_c values were found by conducting simulations under the Alt/ML and Alt/MAP 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). A range of possible z_c values was considered; this range extended from $z_c = 0.6$ to $z_c = 1.4$ in increments of 0.1. Thus, a total of nine possible z_c values were tested. Maximum FI item selection was used for all simulations. The item pool used for these simulations was the same as that used for the full study.

For a given test value of z_c and ability estimation procedure, a simulation was conducted with 500 replications per true ability level $\theta = \{-2, -1, 0, 1, 2\}$, such that the total number of replications per simulation was 2500. Each simulation used the item parameters listed in Appendix A, and the maximum test length for each simulation was 25 items. The first set of simulations examined the accuracy of the $\hat{\theta}^*$ alternative ability estimates with respect to examinee true ability. Accuracy in this sense refers to whether the absolute difference from true ability for the alternative procedure, or $\left|\hat{\theta}^* - \theta\right|$, is less than the absolute difference for the conventional procedure, or $\left|\hat{\theta} - \theta\right|$, where $\hat{\theta}$ is either a conventional ML or MAP ability estimate. Note that these accuracy measures may only be obtained for those situations where the alternative ability estimation procedure is invoked.

In Appendix B, the probability that the alternative estimation procedure yields a more accurate ability estimate than the conventional estimate, given that the procedure is invoked, is denoted by P(acc|invoked). (Note that this measure is collected over all true ability levels θ, and so is not conditional on ability.) For a given test value of z_c and ability estimation procedure (Alt/ML or Alt/MAP), this accuracy measure P(acc|invoked) is provided for each item administration number $i = \{1, 2, \ldots, 25\}$. On the basis of these accuracy measures, it was concluded that for the Alt/ML procedure, z_c should be no less than 0.9; for the Alt/MAP procedure, z_c should be no less than 1.1. (See Appendix B for more detail.) Because larger values of z_c necessarily restrict the number of times the alternative procedure is invoked, smaller z_c values that lead to reasonable accuracy measures are preferred.

Further evidence for selecting an optimal z_c value was obtained from the second outcome measure, 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). If the alternative ability estimation procedure is more efficient than the conventional estimation procedure, relative efficiency measures should be greater than 1; conversely, if the alternative procedure is less efficient, the measures will be less than 1. As with the accuracy measures, potential z_c values ranged from 0.6 to 1.4 in steps of 0.1. Simulations were conducted for 2500 examinees (500 per true ability level) and for tests of length 5, 10, 15, and 25 items. Relative efficiency at true ability θ was computed as the ratio of test information at $i = \{5, 10, 15, 25\}$ items under the alternative procedure, or $I_{ALT}^{(T)}(\theta)$, to the test information at i items under the conventional procedure, or $I_{CONV}^{(T)}(\theta)$. Thus, simulations under ML/Alt and ML were used to identify the optimal z_c for the Alt/ML procedure; likewise, simulations under MAP/Alt and MAP were used to identify the optimal z_c for the Alt/MAP procedure. According to the relative efficiency measures, it was concluded that the optimal z_c value for Alt/ML was 0.9; for the Alt/MAP procedure, the optimal z_c value was 1.3. (See Appendix B for more detail.)

Analysis of the accuracy measures for the two procedures suggests than the optimal z_c value for Alt/ML is no less than 0.9, and no less than 1.1 for Alt/MAP, with smaller values of z_c preferred as long as accuracy is maintained. Analysis of the relative efficiency measures converges with the analysis of the accuracy measures, with a recommendation of $z_c = 0.9$ for the Alt/ML procedure and $z_c = 1.3$ for the Alt/MAP procedure. Thus, for this study, $z_c = 0.9$ will be used for the hypothesis tests in the Alt/ML procedure, and $z_c = 1.3$ for the hypothesis tests in the Alt/MAP procedure.

Related methods

The hypothesis-testing approach has been adopted by researchers examining related problems in measurement and testing, especially those exploring person-fit statistics. In particular, the test statistic utilized in the proposed alternative item selection procedure bears some resemblance to those developed for person-fit statistics, such as Trabin & Weiss' (1983) person response function (PRF) chi-square statistic, Tatsuoka's (1984) standardized extended caution index (ECI4$_z$), and Drasgow, Levine, & Williams' (1985) standardized likelihood-based statistic l_z.

Although a brief review of these studies is to follow, it is important at this stage to reiterate the theoretical assumptions underlying the hypothesis-testing procedure used here. These assumptions clearly distinguish the interpretation of the hypothesis tests in the present study from those in the related literature. First, and most importantly, all hypothesis tests in the related studies are dependent on ability estimates[7], whereas the hypothesis tests employed here require no ability estimates in order to function. This lack of dependence is a direct consequence of the formulation of the null hypothesis in this study. The null hypothesis is stated under the assumption of an ideally-functioning CAT, the behavior of which is wholly predicted by the IRT model in a special case where measures of ability need not be known. That is, if a CAT is administering items targeted at an examinee's ability, then it must be the case that each item obtains maximum information at that examinee's ability. The expected proportion correct for each of these items, given the assumption of perfect targeting, is thus obtainable from the item parameters alone. The hypothesis-testing procedure used here then compares the observed proportion correct to the expected proportion correct, and a decision rule is followed. At no point is an estimate of examinee ability employed in the hypothesis-testing procedure.

The hypothesis-testing procedures considered next, all drawn from the related literature, utilize estimates of ability. The similarities of the test statistics employed in these studies to the test statistic used here are worthwhile to examine from a mathematical standpoint; however, it must be stressed that the assumptions underlying each of these procedures are not identical. Statistics developed for person-fit, as outlined by Nering & Meijer (1998) and Nering (1997), are considered next.

Person-fit statistics. The motivation for person-fit statistics is "to identify response patterns that are incongruent with the underlying test model" (Nering, 1997).

[7] Typically, ability estimates are used. However, in simulation studies the true parameters θ may be examined. Nevertheless, some measure of ability is required for the computation of these statistics.

To this end, a number of person-fit statistics have been developed. Of the three considered here, the earliest proposed was Trabin & Weiss' (1983) person response function chi-square statistic, followed by Tatsuoka's (1984) extended caution index (ECI4*z*), and finally Drasgow et al.'s (1985) l_z statistic.

Trabin & Weiss (1983) propose a chi-square statistic to test the fit of what they refer to as the person response function (PRF) to the data. Their chi-square test is formulated in a similar manner as those used to assess model-data fit for logistic regression applications: (1) items are ordered according to their difficulty parameters; (2) *G* strata of items are formed, where each of the $g = 1, 2, \ldots, G$ strata contain *K* items, with *K* constant across strata; (3) chi-square terms are computed for the difference between the observed number correct in each stratum $\sum_{k=1}^{K} X_k$ and the expected number correct $\sum_{k=1}^{K} P_k\left(\hat{\theta}\right)$, where $\hat{\theta}$ is an ability estimate; and (4) the sum of these chi-square terms is taken over all *G* strata. Under the null hypothesis, where it is assumed that the PRF is consistent with the underlying IRT model, the final statistic is distributed as χ^2 with *G*-2 degrees of freedom in the case of dichotomous item responses. The Trabin & Weiss χ^2 statistic is then defined as

$$\chi^2 = \sum_{g=1}^{G} \left\{ \frac{\left[\sum_k X_k - \sum_k P_k\left(\hat{\theta}\right) \right]^2}{\sum_k P_k\left(\hat{\theta}\right)} + \frac{\left[\sum_k (1 - X_k) - \sum_k \left(1 - P_k\left(\hat{\theta}\right)\right) \right]^2}{\sum_k \left(1 - P_k\left(\hat{\theta}\right)\right)} \right\} \qquad \text{(Eq. 45)}$$

which, when simplified, yields

$$\chi^2 = \sum_{g=1}^{G} \left\{ \frac{\left[\sum_k X_k - \sum_k P_k\left(\hat{\theta}\right)\right]^2}{\sum_k P_k\left(\hat{\theta}\right)} + \frac{\left[\sum_k P_k\left(\hat{\theta}\right) - \sum_k X_k\right]^2}{k - \sum_k P_k\left(\hat{\theta}\right)} \right\} \qquad \text{(Eq.46)}$$

This procedure assumes that for any stratum *g*, the expected proportions correct $P_k\left(\hat{\theta}\right)$ should be similar across the *K* items within that stratum. With this assumption, it can be shown that the χ^2 statistic is equivalent to a sum of squared random normal variables z_g^2, such that $\chi^2 = \sum_{g=1}^{G} z_g^2$, with z_g given by

$$z_g = \frac{\sum_k \frac{X_k}{k} - \sum_k \frac{P_k\left(\hat{\theta}\right)}{k}}{\sqrt{\dfrac{\sum_k \frac{P_k\left(\hat{\theta}\right)}{k}\left(1 - \sum_k \frac{P_k\left(\hat{\theta}\right)}{k}\right)}{k}}} \qquad \text{(Eq.47)}$$

If $P_k\left(\hat{\theta}\right) \approx p$ for all *k*, then Equation 47 reduces to the familiar expression

$$z_g = \frac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{k}}} \qquad \text{(Eq.48)}$$

where $\hat{p} = \dfrac{\sum_k X_k}{k}$. This equation bears resemblance to the test statistic z^* employed in the alternative item selection procedure (see Equation 39). However, the z^* statistic does not assume that the expected proportions $P_k\left(\hat{\theta}\right)$ be nearly equivalent over all items *k*.

To show that $\chi^2 = \sum_{g=1}^{G} z_g^2$, begin with the g^{th} term in Equation 46, and simplify notation by letting the observed number of correct responses $\sum_k X_k$ be denoted by O_g

and the expected number of correct responses $\sum_k P_k\left(\hat{\theta}\right)$ be denoted by E_g. Then the g^{th} term in Equation 46 becomes

$$X_g^2 = \frac{\left[O_g - E_g\right]^2}{E_g} + \frac{\left[E_g - O_g\right]^2}{k - E_g}$$

which may be expressed as

$$X_g^2 = \frac{\left(k - E_g\right)\left[O_g - E_g\right]^2 + E_g\left[E_g - O_g\right]^2}{E_g\left(k - E_g\right)}$$

This expression simplifies to

$$X_g^2 = \frac{k\left(O_g - E_g\right)^2}{kE_g - E_g^2} \qquad \text{(Eq. 49)}$$

Now, if the same substitution of O_g and E_g is applied to Equation 47, the following is obtained

$$z_g = \frac{\frac{O_g}{k} - \frac{E_g}{k}}{\sqrt{\frac{\frac{E_g}{k}\left(1 - \frac{E_g}{k}\right)}{k}}} = \frac{\frac{1}{k}\left(O_g - E_g\right)}{\sqrt{\frac{\frac{E_g}{k} - \frac{E_g^2}{k^2}}{k}}} \qquad \text{(Eq. 50)}$$

After squaring Equation 50,

$$z_g^2 = \frac{\frac{1}{k^2}\left(O_g - E_g\right)^2}{\left[\frac{E_g}{k^2} - \frac{E_g^2}{k^3}\right]} = \frac{k\left(O_g - E_g\right)^2}{kE_g - E_g^2} \qquad \text{(Eq. 51)}$$

Both Equations 49 and 51 are thus shown equivalent, therefore $\chi^2 = \sum_g X_g^2 = \sum_g z_g^2$.

Although a mathematical relationship between the Trabin & Weiss χ^2 statistic and the z^* statistic used in the alternative item selection procedure has been shown, there are clear differences. First, computation of the χ^2 test statistic requires a partitioning of administered items into relatively homogeneous subsets, while the *z*-test in Equation 39 does not. However, even if partitioning were performed, Equation 39 accounts for variation in the expected proportions correct differently than the partitioned z_g statistic of Equation 47. In the case of Equation 47, the standard error of the expected proportions is computed as a function of the mean proportion, such that the standard error is taken as $\sqrt{\frac{\bar{p}(1-\bar{p})}{k}}$ for *k* items. For Equation 39, the standard error is $\sqrt{\frac{\sum p_k(1-p_k)}{k^2}}$, where p_k is the expected proportion correct for item *k*.

The next two person-fit statistics considered are Tatsuoka's (1984) extended caution index (ECI4$_z$) and Drasgow et al.'s (1985) l_z index. Both indices rely on the normal distribution for hypothesis testing; in this way, they are similar to the test statistic used in the present study. Tatsuoka's ECI4$_z$ index is quite similar to the z^* statistic of Equation 39, and would in fact be identical (within a change of sign) if the weights $W_k = \left(p_k - \frac{1}{k}\sum p_k\right)$ in the ECI4$_z$ index were not present. The ECI4$_z$ index is defined as

$$\text{ECI4}_z = \frac{\sum_{k=1}^{n} p_k W_k - \sum_{k=1}^{n} X_k W_k}{\left[\sum_{k=1}^{n} p_k (1-p_k) W_k^2\right]^{1/2}} \qquad \text{(Eq. 52)}$$

where $p_k = P_k(\hat{\theta})$ is the expected proportion correct for item *k*, X_k is the observed response for item *k*, and W_k is the weight for item *k* as defined previously. Note that if the weights were not present, the index would equal the negative of z^*, since

$$\frac{\sum_{k=1}^{n} p_k - \sum_{k=1}^{n} X_k}{\left[\sum_{k=1}^{n} p_k\left(1-p_k\right)\right]^{1/2}} = \frac{\frac{1}{k}\sum_{k=1}^{n} p_k - \frac{1}{k}\sum_{k=1}^{n} X_k}{\left[\frac{1}{k^2}\sum_{k=1}^{n} p_k\left(1-p_k\right)\right]^{1/2}} = \frac{\frac{1}{k}\sum_{k=1}^{n} p_k - \hat{p}}{\left[\frac{1}{k^2}\sum_{k=1}^{n} p_k\left(1-p_k\right)\right]^{1/2}} = -z^* \qquad \text{(Eq. 53)}$$

Once the ECI4$_z$ index for a person's response pattern is obtained, it may be compared to the standard normal distribution. Those indices lying in the rejection region of the $N(0,1)$ distribution suggest that the IRT model is not consistent with the person's response pattern.

Like Tatsuoka's ECI4$_z$, Drasgow et al.'s (1985) l_z index is also compared to the standard normal distribution; however, the construction of the index is based on observed and expected likelihood functions rather than on observed and expected proportions correct. In this way, l_z is formulated quite differently than Trabin & Weiss' χ^2, Tatsuoka's ECI4$_z$, and the z-statistic employed in the present study. Drasgow et al. define the l_z statistic as

$$l_z = \frac{l_0 - E\left[l_0\right]}{\sqrt{Var\left[l_0\right]}} \qquad \text{(Eq. 54)}$$

where the observed likelihood function l_0 is given by

$$l_0 = \sum_{i=1}^{n} u_i \ln P_i\left(\hat{\theta}\right) + \left(1-u_i\right)\ln\left[1-P_i\left(\hat{\theta}\right)\right] \qquad \text{(Eq. 55)}$$

for observed dichotomous response $u_i = \{0,1\}$. The expected value $E\left[l_0\right]$ is given by

$$E\left[l_0\right] = \sum_{i=1}^{n} P_i\left(\hat{\theta}\right)\ln P_i\left(\hat{\theta}\right) + \left[1-P_i\left(\hat{\theta}\right)\right]\ln\left[1-P_i\left(\hat{\theta}\right)\right] \qquad \text{(Eq. 56)}$$

and the variance $Var[l_0]$ by

$$Var[l_0] = \sum_{i=1}^{n} P_i(\hat{\theta})[1 - P_i(\hat{\theta})]\left\{\ln\left\{\frac{P_i(\hat{\theta})}{[1 - P_i(\hat{\theta})]}\right\}\right\}^2 \qquad \text{(Eq. 57)}$$

Experimental design

The four factors in the experimental design were: (1) item selection procedure (maximum FI item selection 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 ($\theta = \{-2,-1,0,+1,+2\}$); and (4) test length (5, 10, 15, or 25 items). The experimental design was a fully-crossed $2 \times 5 \times 5 \times 4$ design, with two levels for item selection procedure, five levels for ability estimation procedure, five levels for examinee true ability, and four levels for test length. (Note that by this design all tests in the study are of fixed length.) For each of the experimental conditions, 1000 replications were generated. That is, simulated response patterns for 1000 subjects were generated for each cell in the design.

Efficiency, as defined by Equation 32, was the primary dependent measure. Since this measure is rather highly skewed to the left, the median efficiency was reported as a measure of central tendency. The interquartile range (that is, the range in the efficiency measure between the 25th and 75th percentile points) was reported as a measure of variability. In addition, efficiency measures at the 25th and 75th percentile points were reported in order to make more detailed comparisons of efficiency across the experimental conditions possible. An example of the distribution of the efficiency measures for a CAT administered to 500 examinees with $\theta = 0$ using maximum FI item selection and ML ability estimation is shown in Figure 2. The distribution of efficiency is provided for tests of length 5, 10, 15, and 25 items.

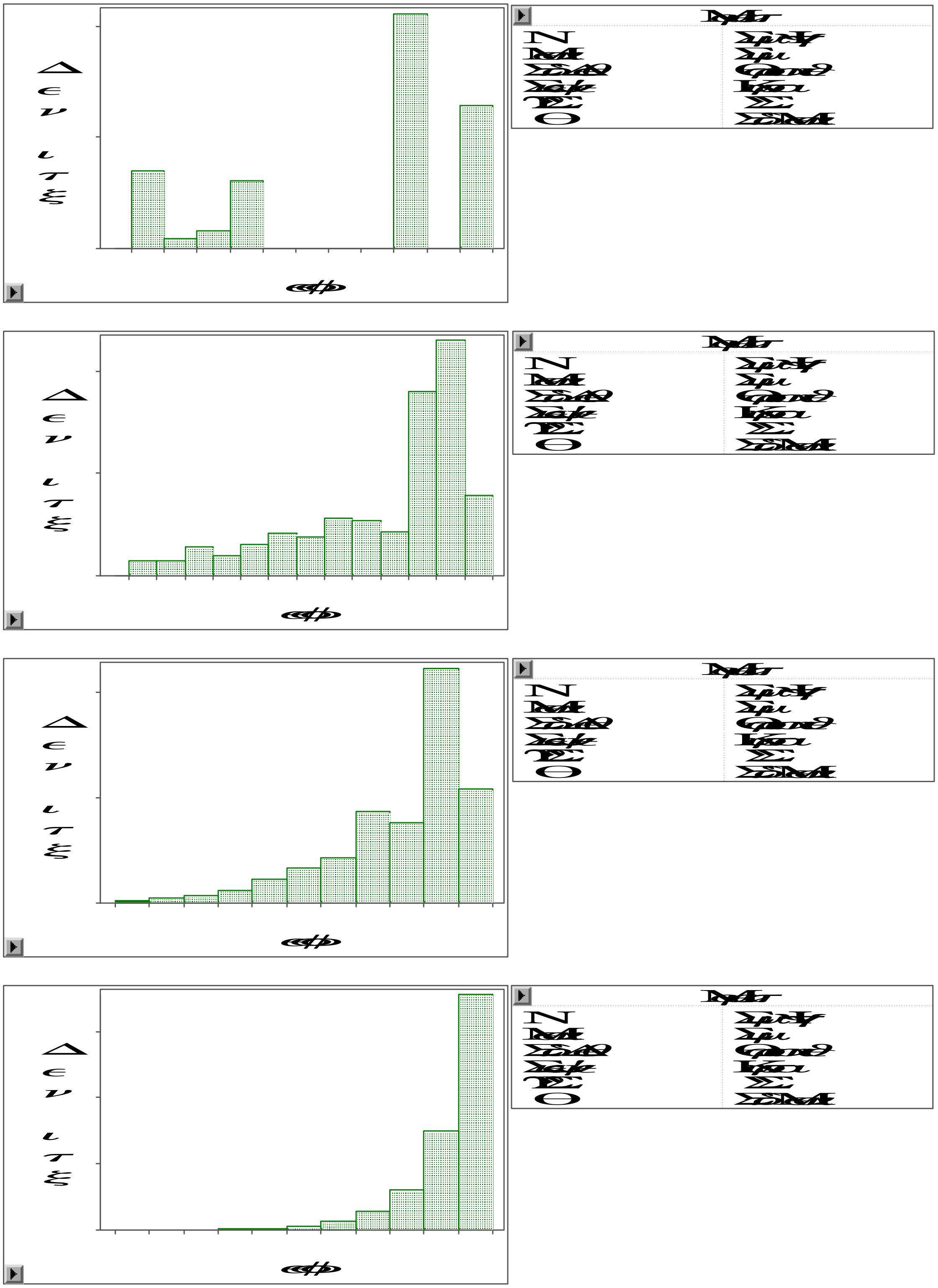

Figure 2. Distribution of efficiency measures for tests of length 5, 10, 15, and 25 items; *N* = 500 examinees with θ = 0, maximum FI item selection, ML ability estimation.

The efficiency measures were calculated according to the accumulated information terms in Equation 32, where

$$Efficiency\left(CAT|\theta\right) = \frac{I_{CAT}^{(T)}}{I_0^{(T)}} = \frac{I_{CAT}^{(T)}}{I_1\left(\theta\right) + \sum_{j=2}^{J} I_j\left(\theta\right)}$$

The numerator of this expression is the accumulated information at examinee true θ from a CAT administration with item selection by FI or FII and ability estimation by ML, MAP, GSS, ML/Alt, or MAP/Alt. The denominator of this expression is accumulated test information from the optimal subset of items, as defined by Equation 31. To obtain the quantity $I_0^{(T)}$, one CAT administration was conducted for each of the test length manipulations in the experimental design, where items were selected according to examinee true θ, as opposed to the ability estimates obtained from ML, MAP, GSS, ML/Alt, or MAP/Alt. One administration is sufficient to obtain $I_0^{(T)}$ since both true ability θ and its corresponding $I_0^{(T)}$ are constants.

Computation of the median efficiency, as well as the efficiency at the 25th and 75th percentile points, follows from Equation 32 and uses the corresponding $I_{CAT}^{(T)}$ measures at each percentile point. Thus, the efficiency measure at the pth percentile point is defined as

$$P_p\left[Efficiency\left(CAT|\theta\right)\right] = \frac{P_p\left(I_{CAT}^{(T)}\right)}{I_0^{(T)}} \qquad \text{(Eq. 58)}$$

where $P_p(\ )$ indicates the value of the measure at the pth percentile. In this manner, the interquartile range (IQR) may be defined as

$$IQR\left[Efficiency\left(CAT\middle|\theta\right)\right]=\frac{P_{75}\left(I_{CAT}^{(T)}\right)-P_{25}\left(I_{CAT}^{(T)}\right)}{I_{0}^{(T)}} \qquad \text{(Eq. 59)}$$

Although the efficiency measures are of primary interest in this study, as a ratio of information measures they cannot, by themselves, indicate the magnitude of accumulated test information. Thus, alongside each set of summary statistics for the efficiency measure (i.e., median, IQR, and 25th and 75th percentile points), the quantity $I_0^{(T)}$ is provided. Hence, any efficiency measure may be converted to an information measure through $I_0^{(T)}$.

In addition to the efficiency measures, the mean and standard deviation of the distribution of provisional ability estimates $\hat{\theta}$ for the examinees within each cell in the design will be computed. Note that for ability estimation by ML/Alt or MAP/Alt, two ability estimates are possible: one from the conventional procedure (ML or MAP), the other from the alternative procedure. The particular estimate (i.e., conventional or alternative) that is used to select the next item is recorded as the provisional ability estimate $\hat{\theta}$ for an examinee when ML/Alt or MAP/Alt is employed as the ability estimation procedure.

The layout of the experimental design is illustrated in Table 5.

Table 5. 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 25th, 50th (median), 75th percentiles, IQR • Mean and S.D. of provisional ability estimates				

To perform the CAT simulations required for the study, an item bank of 367 pre-calibrated and dichotomously-scored 3P items from a recently-administered large-scale CAT assessment of mathematics ability was used. In the logistic metric where the scaling parameter $D = 1.7$, the mean and standard deviation of the discrimination parameters (i.e., *a* parameters) from the 367 items are 0.950 and 0.341, respectively. For the difficulty parameters (i.e., *b* parameters), the mean and standard deviation are 0.158 and 1.113, respectively. For the pseudo-guessing parameters (i.e., *c* parameters), the mean and standard deviation are 0.144 and 0.105, respectively. In its operational form, the CAT administered using this item bank is fixed at a length of 28 items; however, as it was hypothesized that the greatest variation in CAT efficiency would occur much earlier (e.g., at or before the 10th administered item), the CAT simulations were fixed such that no test exceeded a length of 25 items. The item parameters from this pool of 367 items may be found in Appendix A.

CAT simulation method

Central to this study is a method for simulating a CAT administration. Simulation is necessary to conduct the study since the efficiency measure as defined in Equation 32 requires an examinee's true ability level θ to be known in advance. Further, controlled comparisons between the efficiency of maximum FI item selection and the proposed alternative item selection procedure can only be made using a simulation design.

Simulated CAT administrations were generated using the program SimCAT, a multi-purpose CAT simulation program written in the SAS language (SAS Institute, 2000). This program was designed to handle all manipulations in the experimental design (i.e., those corresponding to the item selection procedure, ability estimation procedure, test length, and true ability at discrete points), as well as to provide the accumulated test information measures used to calculate efficiency.

As simulation techniques may vary from program to program, a brief review of the general operating characteristics of SimCAT is worthwhile. First, pseudo-random numbers used by the program are generated from a starting seed value, which is pre-loaded into the program from a user-specified file before any response vectors are generated. That is, the sequence of pseudo-random numbers used to generate item responses is reproducible, as the seed value is fixed as opposed to being dynamic (as would be the case if the internal clock value was used to generate the random number seed). Second, SimCAT generates item responses for all items in the item pool prior to CAT administration, even if those items are not selected by the CAT algorithm for administration to the examinee. All examinee responses are generated according to the

3P IRT model, with the following rule applied for determining whether a 0 (incorrect response) or 1 (correct response) should be assigned. Given an examinee with true ability θ and an item *i* with parameter vector $\omega_i = \{a_i, b_i, c_i\}$, the probability of correct response is given by $P_i(U_i = 1|\theta, \omega_i)$; if a random *UNIF*(0,1) deviate r_i is obtained where $r_i \le P_i(U_i = 1|\theta, \omega_i)$, then the observed response $X_i = 1$, otherwise $X_i = 0$.

Ability estimation in SimCAT for both MAP and ML estimation is by Newton-Raphson, with Fisher scoring used for the second derivative of the log-likelihood function. In cases where the second derivative is not negative definite, a grid search for the maximum of the likelihood function is performed to obtain the ability estimate $\hat{\theta}$. For MAP ability estimation, the informative prior is assigned to be *N*(0,1). The asymptotic variance of the ML estimates is computed according to Equation 13; for MAP estimates, Equation 14 is used. All CAT simulations begin with an initial estimate of examinee ability $\hat{\theta} = 0$, and so for any given item selection procedure, all examinees receive the same first item. Thus, the modification to the efficiency measure as shown in Equation 31 is appropriate.

For maximum FI item selection, SimCAT searches the item pool for the next available item whose information is maximum at the provisional ability estimate $\hat{\theta}$. SimCAT evaluates the information function for all potential items *i* in the pool at $\hat{\theta}$, as opposed to using an information lookup table. Thus, SimCAT evaluates $I_i(\hat{\theta})$ for all *i* and selects the item whose $I_i(\hat{\theta})$ is maximum. Once an item has been administered, it may no longer be considered for subsequent administration. For maximum FII item selection, a 95% modified confidence interval about $\hat{\theta}$ is generated (as suggested by

Chen, Ankenmann & Chang, 2000; p. 248), such that the lower bound θ_l and upper bound θ_u are given by

$$\theta_l = \hat{\theta} - \frac{1.96}{\sqrt{n+1}}, \quad \theta_u = \hat{\theta} + \frac{1.96}{\sqrt{n+1}} \qquad \text{(Eq. 60)}$$

where n is the number of items administered. The FII for an item i is defined as

$$FII_i(\theta) = \int_{\theta_l}^{\theta_u} I_i(\theta) d\theta \qquad \text{(Eq. 61)}$$

SimCAT uses the exact solution for the integral given by Equation 61. Because the solution is rather lengthy, it is provided in Appendix D.

The ability scale used in the simulations imposes a negative bound at −4 and a positive bound at +4. These bounds are utilized both by the information table and the ability estimation routines. In cases where ML estimates are undefined (i.e., a response pattern of all 0's or all 1's), the lower bound (for a vector of all incorrect responses) or upper bound (for a vector of all correct responses) on the ability scale is taken as the estimate.

CHAPTER 4

Results

This chapter presents the results of the study discussed in the methodology section. Recall that the four factors in the experimental design were: (1) item selection procedure (maximum FI item selection or maximum FII item selection); (2) ability estimation procedure (ML, ML/Alt, MAP, MAP/Alt, or GSS); (3) true ability level at discrete points along the ability continuum ($\theta = \{-2,-1,0,+1,+2\}$); and (4) test length (5, 10, 15, or 25 items). For each of the experimental conditions, 1000 replications were generated using a simulation methodology.

The primary dependent measure in the study was efficiency, as defined by Equation 32. To facilitate the reporting of the results, the proportions calculated by Equation 32 were converted to percentages; these percentages are given in the tables and figures that follow. Results are summarized in terms of median efficiency, the interquartile range of efficiency, and the 25th and 75th percentile points of efficiency for each of the experimental conditions. Efficiency measures may be converted to information measures using Equation 32 and the accumulated test information measures $I_0^{(T)}(\theta)$ provided for an optimal subset of items given true ability and test length. In addition, the means and standard deviations of the provisional ability estimates are provided.

In order to facilitate the comparisons across the experimental conditions as indicated by the research questions, the results are presented both in tabular and graphical format and organized in the following manner. First, results from item selection under maximum Fisher information (FI) are tabled separately from those obtained under maximum Fisher interval information (FII). Within a level of item selection procedure, three tables are provided for each combination of ability estimation procedure × true ability level × test length: (1) medians and interquartile ranges (IQRs) of the efficiency measure; (2) 25th and 75th percentile points of the efficiency measure; and (3) means and standard deviations of the provisional ability estimates. A supplementary table provides the accumulated test information measures $I_0^{(T)}(\theta)$ for true ability level × test length.

Graphical presentation of the results is also first categorized according to item selection procedure (i.e., maximum FI or maximum FII). However, results are further subdivided according to test length (i.e., 5, 10, 15, or 25 items). Thus, each figure presents dependent measures for each combination of ability estimation procedure × true ability level for the specified levels of item selection procedure and test length. The dependent measures displayed graphically are median efficiency and efficiency IQR.

Analysis of efficiency measures

Maximum FI item selection

The efficiency measures from maximum FI item selection are summarized in Tables 6 and 7. Both tables summarize the efficiency measures for the experimental conditions of ability estimation procedure × test length × true ability level. Table 6 provides the medians and interquartile ranges (IQRs) of the efficiency measures, and

Table 7 provides the 25th and 75th percentile points of the efficiency measures. As discussed in the methodology section, the efficiency measure is skewed to the left; thus, medians are reported as a measure of central tendency and IQRs as a measure of variability. All efficiency measures are by definition bounded below by 0% and above by 100%. Table 8 provides the accumulated test information measures $I_0^{(T)}(\theta)$ for an optimal subset of items; this supplementary information may be used for converting the efficiency measures to accumulated test information measures.

Table 6. Medians and IQRs 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 7. Efficiency measures at the 25th and 75th percentile points under maximum FI item selection.

		Ability level									
		θ = -2		θ = -1		θ = 0		θ = 1		θ = 2	
Ability estimation	Test length	P_{25}	P_{75}	P_{25}	P_{75}	P_{25}	P_{75}	P_{25}	P_{75}	P_{25}	P_{75}
ML	5	36.8	81.0	69.7	81.8	35.4	64.8	33.4	78.8	47.7	64.8
	10	72.5	91.3	70.7	89.5	50.2	79.1	49.5	84.6	73.4	86.6
	15	87.8	96.9	77.6	93.4	64.8	86.6	67.9	89.9	87.0	94.0
	25	93.4	98.9	86.6	96.3	81.4	93.6	81.9	93.5	93.7	97.0
ML/Alt	5	81.5	100.0	88.3	96.1	44.9	71.6	46.8	86.8	80.3	96.8
	10	88.3	100.0	81.1	97.6	58.3	85.3	63.7	94.3	86.2	100.0
	15	95.9	99.9	84.1	98.4	71.0	92.5	75.2	95.8	94.9	99.9
	25	96.7	99.0	90.2	98.7	84.3	96.9	87.2	98.5	97.8	100.0
MAP	5	8.1	57.3	77.5	97.1	77.2	92.9	76.4	100.0	23.6	23.6
	10	52.7	81.8	77.5	97.7	81.4	96.2	83.3	96.9	59.2	72.8
	15	73.0	91.4	82.6	98.7	85.4	97.8	86.9	98.0	78.2	87.3
	25	86.3	94.4	89.8	98.9	92.9	98.8	91.5	98.8	91.3	94.4
MAP/Alt	5	57.3	79.7	77.5	92.4	71.6	92.9	68.4	90.6	54.6	54.6
	10	68.8	90.6	77.5	96.7	79.2	93.7	77.4	94.5	63.2	82.8
	15	76.4	95.5	83.4	98.1	85.4	97.6	83.5	97.4	81.7	91.8
	25	88.7	97.4	91.1	98.9	92.1	98.8	90.5	98.8	92.2	95.4
GSS	5	79.7	96.7	67.0	87.9	63.3	83.1	51.3	83.0	78.3	81.1
	10	83.9	99.6	82.3	94.0	62.4	88.3	61.5	92.5	82.4	94.7
	15	91.2	98.4	83.8	96.0	71.1	91.4	75.7	94.1	91.9	98.2
	25	93.4	98.9	88.2	97.9	84.0	96.1	87.1	97.2	96.1	99.1

Table 8. Accumulated test information measures $I_0^{(T)}(\theta)$ for an optimal subset of items under maximum FI item selection.

	$I_0^{(T)}(\theta)$				
Test length	θ=-2	θ=-1	θ=0	θ=1	θ=2
5	2.351	2.817	5.282	6.461	6.379
10	4.697	5.478	9.571	12.859	12.237
15	6.457	7.862	13.126	18.558	16.495
25	9.351	11.85	18.581	28.149	23.562

Comparison of the efficiency measures is facilitated by Figures 3 through 10. For each level of test length (5, 10, 15, or 25 items), two figures are provided: the first reports the median efficiency, and the second the efficiency interquartile range. True ability level is categorized by the cluster of five bars within an ability estimation procedure; the bars correspond to $\theta = \{-2, -1, 0, 1, 2\}$ from left to right.

Test length of 5 items. Median efficiency measures and interquartile ranges (IQRs) of the efficiency measures for tests of 5 items in length are displayed in Figures 3 and 4. With respect to median efficiency for the conventional ability estimation procedures (ML and MAP), maximum FI item selection performed differently across the range of ability levels depending on whether ML or MAP estimation was used. MAP was clearly superior to ML for $\theta = \{-1, 0, 1\}$; however, ML was more efficient than MAP at the extreme ability levels $\theta = \{-2, 2\}$. MAP achieved a median efficiency exceeding 88% for $\theta = \{-1, 0, 1\}$, but performed poorly at the extremes, with median efficiencies of 31% and 24% at $\theta = \{-2, 2\}$, respectively. Although ML did not suffer the dramatic drop in efficiency at the extremes, it was less efficient than MAP at the middle of the ability continuum, having achieved a maximum median efficiency of 73% at $\theta = -1$ and minimum median efficiency of 45% at $\theta = 1$. Median efficiencies for ML at $\theta = \{-2, 2\}$ were 53% and 62%, respectively, and were higher than those observed for MAP.

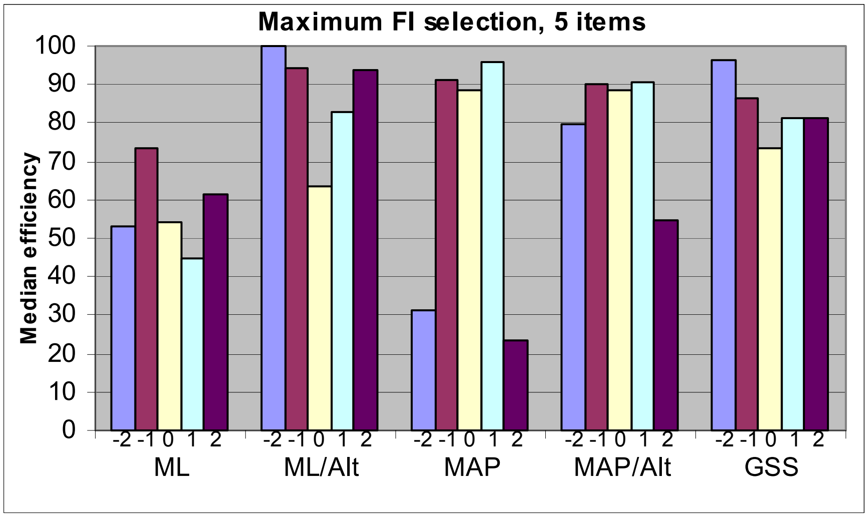

Figure 3. Median efficiency measures under maximum FI item selection for a test of 5 items.

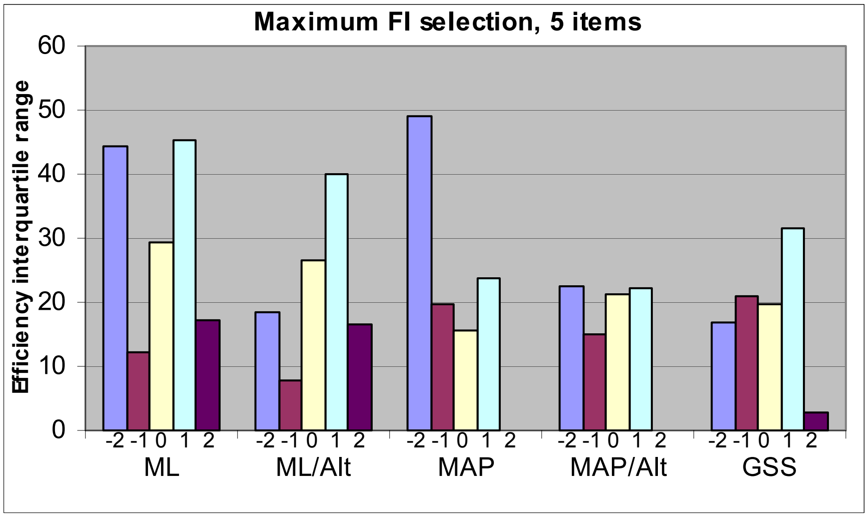

Figure 4. IQRs of the efficiency measure under maximum FI item selection for a test of 5 items. (Note: At θ = 2, the IQRs for MAP and MAP/Alt are equal to zero.)

At this test length, the alternative ability estimation procedures performed better overall than the conventional ability estimation procedures. Median efficiencies for ML concurrent with the alternative ability estimation procedure (ML/Alt) were greater than the corresponding measures observed for conventional ML, and strongly so for $\theta = \{-2, -1, 1, 2\}$. For these ability levels, ML/Alt achieved a maximum median efficiency of 100% at $\theta = -2$ and a minimum median efficiency of 63% at $\theta = 0$. For MAP concurrent with the alternative ability estimation procedure (MAP/Alt), the median efficiency measures were comparable to those observed for conventional MAP at $\theta = \{-1, 0, 1\}$, but median efficiency at the extremes was higher, at 80% and 55% for $\theta = \{-2, 2\}$, respectively. GSS performed similarly to ML/Alt, with slightly lower median efficiency measures than ML/Alt at $\theta = \{-2, -1, 1, 2\}$, but with a higher median efficiency measure of 74% at $\theta = 0$.

Variability in the efficiency measures for the conventional ability estimation procedures, as indicated by IQR, was less for MAP than ML except at $\theta = \{-2, -1\}$. At $\theta > -1$, the IQRs for MAP were at least 10% less than the respective IQRs for ML, with the IQR for MAP at $\theta = 2$ dropping to zero. For ML and $\theta > -1$, efficiency was least variable at $\theta = 2$ and most variable at $\theta = 1$; the respective IQRs at these points were 17% and 45%. At $\theta = -1$, the IQR for MAP was higher than for ML, at 20% versus 12%. At the lowest extreme of ability, $\theta = -2$, the IQR of efficiency for both ML and MAP was approximately 44% and 49%, respectively. If the 25th and 75th percentile points are considered at this ability level, then ML efficiency ranged from 37% to 81%, while MAP efficiency ranged from 8% to 57%. At the other extreme of ability, $\theta = 2$, ML also outperformed MAP, even though the IQR for MAP was 0%. The range in ML efficiency at $\theta = 2$ was 48% to 65%, while it was constant at 24% for MAP. For the middle range of ability, $-2 < \theta < 2$, MAP clearly outperformed ML.

It was observed that the alternative ability estimation procedures achieved higher median efficiencies than the conventional ability estimation procedures. In general, the efficiency measures from the alternative procedures were also no more variable, and in a number of cases less variable, than the conventional procedures. Reduction in variability was rather dramatic for ML/Alt and MAP/Alt at $\theta = -2$; for ML/Alt, the IQR was 19% in contrast to 44% for ML. Likewise, for MAP/Alt the IQR was 23%, in contrast to 49% for MAP. At $\theta = 2$, the IQR for MAP/Alt was equal to zero, a result also observed for conventional MAP. Variability in GSS was similar to that observed for MAP/Alt at $\theta = \{-2, 0, 2\}$, but with somewhat higher IQRs of 21% and 32% at $\theta = \{-1, 1\}$, respectively.

Test length of 10 items. Median efficiency measures and IQRs of the efficiency measures for tests of 10 items in length are displayed in Figures 5 and 6. As observed for tests of 5 items in length, the conventional ability estimation procedures (ML and MAP) performed differently across the range of ability levels in terms of median efficiency. Again, MAP was superior to ML for $\theta = \{-1, 0, 1\}$ and ML was more efficient than MAP at the extreme ability levels $\theta = \{-2, 2\}$. MAP achieved a median efficiency at or above 92% for $\theta = \{-1, 0, 1\}$, but continued to lag in performance at the extremes, with median efficiencies of 73% and 64% at $\theta = \{-2, 2\}$, respectively.

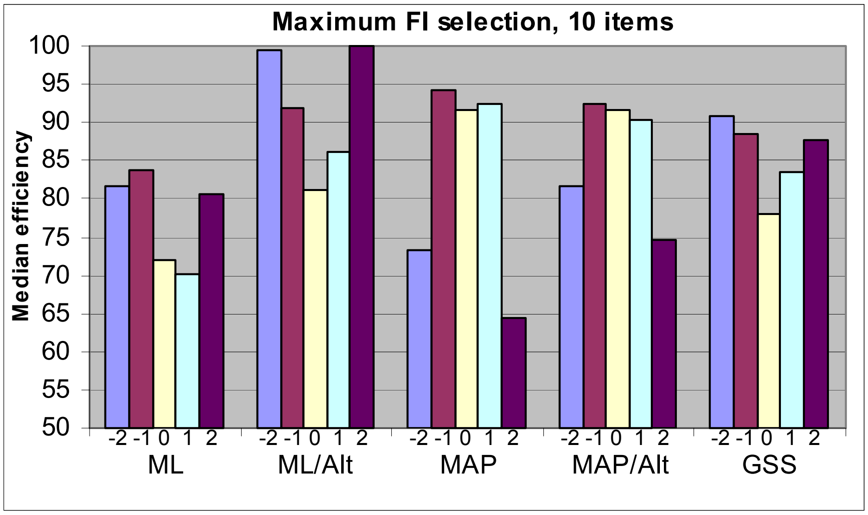

Figure 5. Median efficiency measures under maximum FI item selection for a test of 10 items.

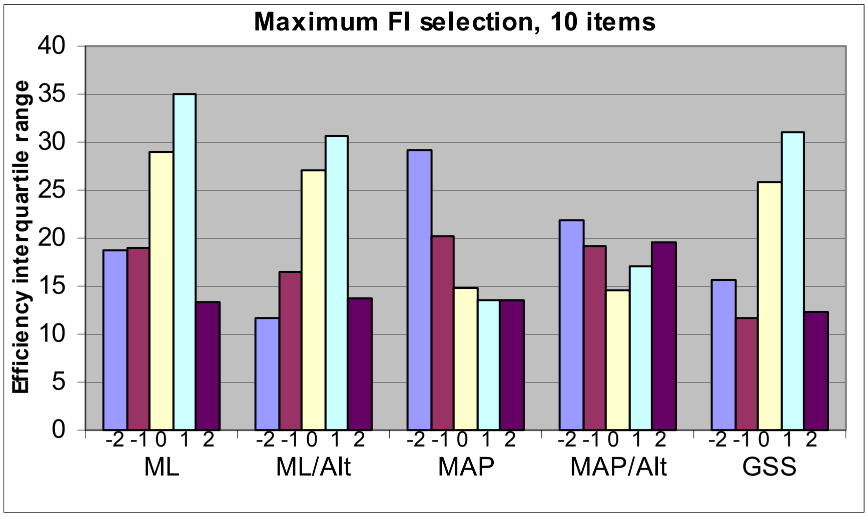

Figure 6. IQRs of the efficiency measure under maximum FI item selection for a test of 10 items.

Nevertheless, the gap in median performance for MAP between the middle ability levels and those at the extremes narrowed from approximately 60% for 5-item tests to approximately 25% for 10-item tests. ML remained less efficient than MAP at the middle of the ability continuum, having achieved a maximum median efficiency of 84% at $\theta = -1$ and minimum median efficiency of 70% at $\theta = 1$. The median efficiency measures for ML at $\theta = \{-2, 2\}$ were again higher than those for MAP, at approximately 81% for these two extreme ability levels.

Like the 5-item tests, the alternative ability estimation procedures performed better overall than the conventional ability estimation procedures for 10-item tests. Median efficiencies for ML/Alt were greater than the corresponding measures observed for conventional ML, with gains of approximately 10% for middle ability levels and 20% for the extreme ability levels. ML/Alt achieved maximum median efficiencies of 100% at $\theta = \{-2, 2\}$, and a minimum median efficiency of 81% at $\theta = 0$. For MAP/Alt, the median efficiency measures were comparable to those observed for conventional MAP at $\theta = \{-1, 0, 1\}$, but median efficiency at the extremes was approximately 10% higher, at 82% and 75% for $\theta = \{-2, 2\}$, respectively. GSS performed similarly to ML/Alt at $\theta = \{-1, 0, 1\}$, but with approximately 10% lower median efficiency measures than ML/Alt at $\theta = \{-2, 2\}$. Median efficiency measures for GSS at $\theta = \{-2, 2\}$ were 91% and 88%, respectively.

Unlike the 5-item tests, variability in the efficiency measures for the conventional ability estimation procedures for tests 10 items in length was substantially less for MAP than ML only at $\theta = 0$ and $\theta = 1$. At ability levels $\theta = \{0, 1\}$, the IQRs for MAP were 15% and 14%, respectively, whereas for ML they were 29% and 35%, respectively. For each ability level $\theta = \{-1, 2\}$, the IQRs for the ML and MAP procedures were nearly identical. For $\theta = -2$, variability in the efficiency measure was less for ML than for MAP, with the respective IQRs being 19% and 29%. If the 25th and 75th percentile points are considered at this ability level, then ML efficiency ranged from 73% to 92%, while MAP efficiency ranged from 53% to 82%. At the other extreme of ability, $\theta = 2$, ML also outperformed MAP. The range in ML efficiency at $\theta = 2$ was 73% to 87%, while it was 59% to 73% for MAP. For the middle range of ability, $-2 < \theta < 2$, MAP clearly outperformed ML.

Variability measures for the alternative procedures were somewhat different in the case of 10-item tests as compared to 5-item tests. Whereas IQRs were much lower at $\theta = -2$ for the alternative procedures than for the conventional procedures on 5-item tests, the reduction in variability was only about 7% for 10-item tests. At $\theta = -2$, the IQR for ML/Alt was 12%, in contrast to 19% for ML. For MAP/Alt, the IQR was 22%, whereas for MAP it was 29%. For ability levels $\theta > -2$, IQRs for ML/Alt were slightly smaller than those for ML; for MAP/Alt, this slight reduction in variability over the conventional procedure was only true for $\theta = \{-1, 0\}$. For $\theta = \{1, 2\}$, the IQRs for MAP/Alt were actually 4% and 6% higher, respectively. The pattern of IQRs for GSS closely mirrored that of ML/Alt, with nearly identical IQRs for $\theta > -1$, a slightly higher IQR at $\theta = -2$, and a slightly lower IQR at $\theta = -1$.

Test length of 15 items. Median efficiency measures and IQRs of the efficiency measures for tests of 15 items in length are displayed in Figures 7 and 8. By this test length, all five ability estimation procedures met or exceeded a median efficiency of 80% across all levels of ability. Further, within each of the conventional ML and MAP procedures, the gaps between maximum and minimum median efficiency measures have narrowed to approximately 10%. For ML, a minimum median efficiency of 80% was obtained at θ = {0, 1}, and a maximum median efficiency of 93% was obtained at θ = −2. For MAP, a minimum median efficiency of 85% was obtained at θ = 2, and a maximum median efficiency of 95% was obtained at θ = 1.

The previous lag in performance observed for MAP at the extremes of ability was nearly absent at 15 items, with only a 5% difference in median efficiency at θ = {−2, 2} between ML and MAP. However, MAP remained superior to ML for θ = {−1, 0, 1}, with a 5% difference in median efficiency between the two procedures at θ = −1, and approximately 14% difference at θ = {0, 1}.

Median efficiency measures from the ML/Alt and MAP/Alt procedures remained higher for the most part than those observed from the respective conventional procedures. ML/Alt achieved greater median efficiency across all ability levels, having maintained a median efficiency of 100% for θ = {−2, 2} and median efficiency measures that exceeded 87% elsewhere. Median efficiencies for ML/Alt at θ = {−2, −1, 0} were approximately 6% higher than the respective measures for ML; at θ > 0, the respective measures were approximately 9% higher. MAP/Alt benefited also at these extreme ability levels, though only by about 4% over MAP. Median efficiencies for MAP/Alt were comparable to those observed for MAP for θ = {−1, 0}, and were slightly less for θ = 1. Median efficiency measures from GSS were similar to ML/Alt, but were approximately 3% less overall.

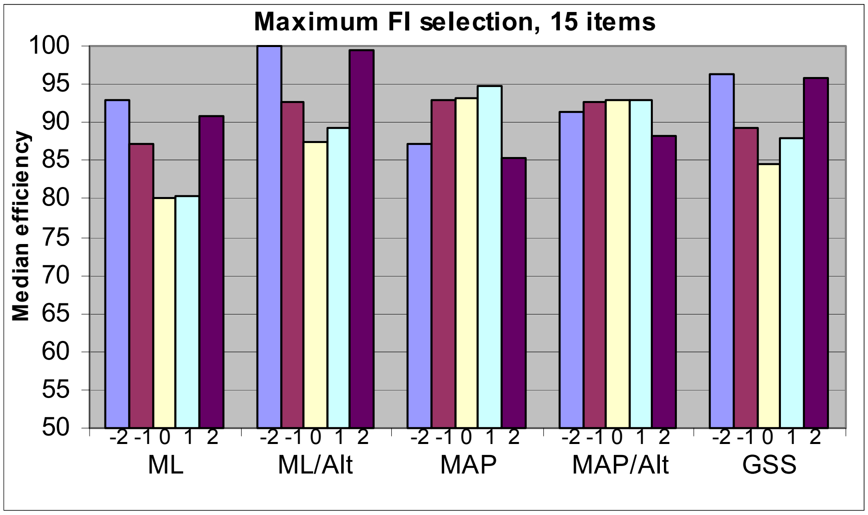

Figure 7. Median efficiency measures under maximum FI item selection for a test of 15 items.

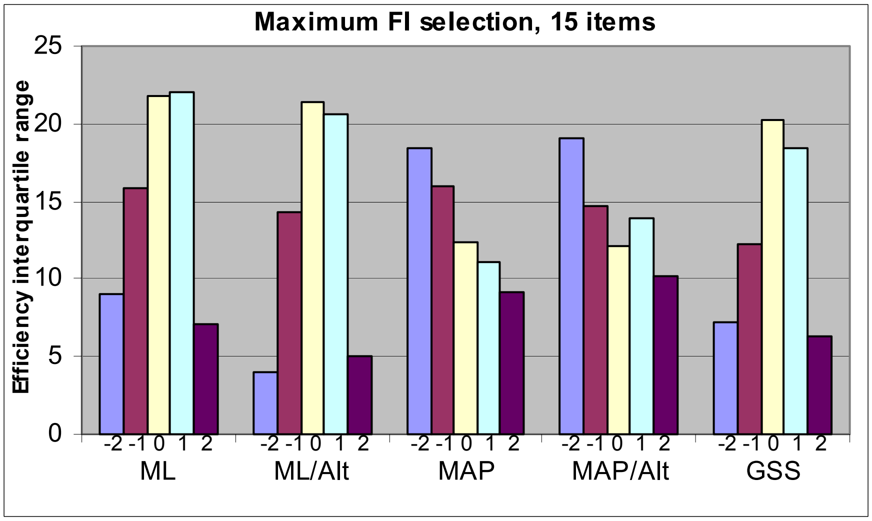

Figure 8. IQRs of the efficiency measure under maximum FI item selection for a test of 15 items.

As observed for the 10-item tests, variability in the efficiency measures for the conventional ability estimation procedures for tests 15 items in length was again less for MAP than ML only at $\theta = 0$ and $\theta = 1$. At these ability levels $\theta = \{0, 1\}$, the IQRs for MAP were 12% and 11%, respectively, whereas for ML they were 22% at both ability levels. For ability levels $\theta = \{-1, 2\}$, the IQRs for the ML and MAP procedures were nearly identical. For $\theta = -2$, variability in the efficiency measure was less for ML than for MAP, with the respective IQRs being 9% and 18%. If the 25th and 75th percentile points are considered at this ability level, then ML efficiency ranged from 88% to 97%, while MAP efficiency ranged from 73% to 91%. At the other extreme of ability, $\theta = 2$, ML also outperformed MAP. The range in ML efficiency at $\theta = 2$ was 87% to 94%, while it was 78% to 87% for MAP. For the middle ability levels $\theta = \{0, 1\}$, MAP clearly outperformed ML, while at $\theta = -1$, the range in ML efficiency was comparable to the range in MAP efficiency.

Variability measures for the ML/Alt and MAP/Alt procedures were comparable to their respective conventional procedures for tests of 15 items in length. While the difference in the variability of efficiency measures between the alternative and conventional procedures for 10-item tests was about 7%, for the 15-item tests the overall difference was negligible. The largest observed difference was 5% at $\theta = -2$ for the ML/Alt procedure, where the IQR was 4% as opposed to 9% for ML. Remaining differences between the alternative and conventional IQRs did not exceed 2%. The pattern of IQRs for GSS again closely mirrored that of ML/Alt, with nearly identical IQRs for $\theta > -1$, a slightly higher IQR at $\theta = -2$, and a slightly lower IQR at $\theta = -1$.

Test length of 25 items. Median efficiency measures and IQRs of the efficiency measures for tests of 25 items in length are displayed in Figures 9 and 10. By this test length, all five ability estimation procedures met or exceeded a median efficiency of 89% across all levels of ability. Further, results at this test length paralleled those observed for the 15-item tests, although median efficiencies were higher overall. Within each of the conventional ML and MAP procedures, the gaps between maximum and minimum median efficiency measures were approximately 7%. For ML, a minimum median efficiency of 89% was obtained at $\theta = \{0, 1\}$, and a maximum median efficiency of 97% was obtained at $\theta = -2$. For MAP, a minimum median efficiency of 91% was obtained at $\theta = 2$, and a maximum median efficiency of 97% was obtained at $\theta = \{0, 1\}$.

Although the pattern of greater performance of MAP at middle ability levels and ML at extreme ability levels in terms of median efficiency measures remained for 25-item tests, the differences in performance were small. The median efficiency for ML at $\theta = -2$ was only 6% higher than the corresponding measure for MAP; likewise, ML efficiency was only 2% higher than MAP at $\theta = 2$. For middle ability levels, the gap between MAP and ML performance narrowed considerably, with MAP having outperformed ML by approximately 7% at $\theta = \{0, 1\}$, and 4% at $\theta = -2$.

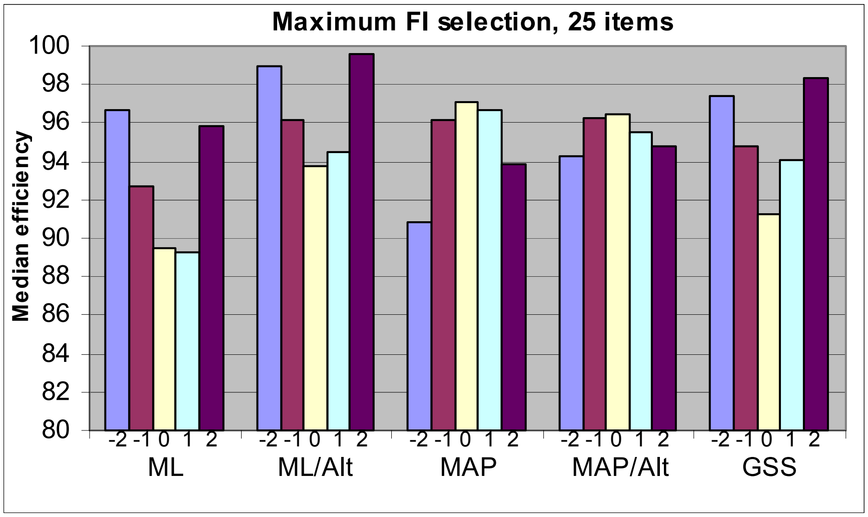

Figure 9. Median efficiency measures under maximum FI item selection for a test of 25 items.

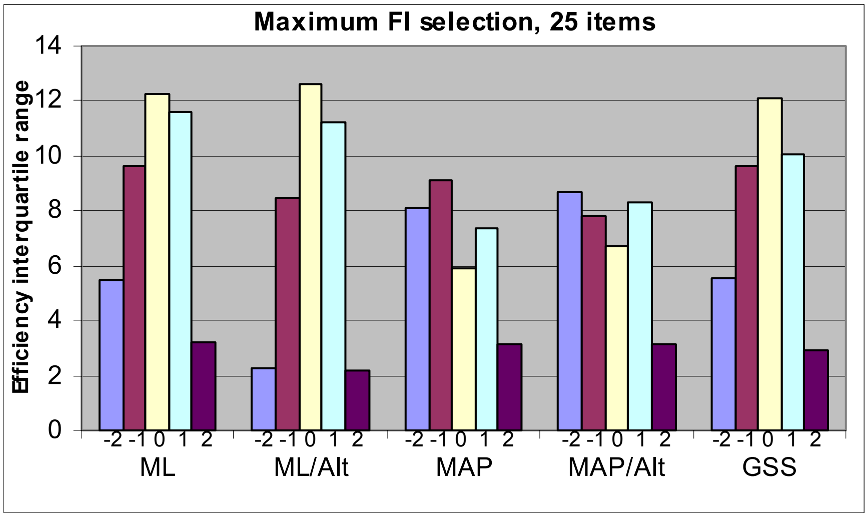

Figure 10. IQRs of the efficiency measure under maximum FI item selection for a test of 25 items.

As observed previously, median efficiency measures from the ML/Alt and MAP/Alt procedures remained higher for the most part than those observed from the respective conventional procedures. However, like the 15-item tests the gains in median efficiency at 25 items were much smaller as compared to those observed for 5- and 10-item tests. ML/Alt again achieved greater median efficiency across all ability levels, with measures of 99% and 100% for $\theta = \{-2, 2\}$, and measures exceeding 94% elsewhere. MAP/Alt benefited only by about 4% over MAP at $\theta = -2$, and elsewhere the measures were comparable, even at $\theta = 2$. Median efficiency measures from GSS were similar to ML/Alt, but were approximately 2% less at $\theta = \{-2, -1, 0\}$ and nearly identical at $\theta > 0$.

Like the 10- and 15-item tests, variability in the efficiency measures for the conventional ability estimation procedures for tests 25 items in length was again less for MAP than ML only at $\theta = 0$ and $\theta = 1$. At these ability levels $\theta = \{0, 1\}$, the IQRs for MAP were 6% and 7%, respectively, whereas for ML they were 12% at both ability levels. For ability levels $\theta = \{-1, 2\}$, the IQRs for the ML and MAP procedures were nearly identical. For $\theta = -2$, variability in the efficiency measure was slightly less for ML than for MAP, with the respective IQRs being 6% and 8%. If the 25th and 75th percentile points are considered, ML outperformed MAP only at $\theta = -2$, where the range in the ML efficiency measures was 93% to 99%, whereas it was 86% to 94% for MAP. For middle ability levels, MAP outperformed ML only at $\theta = \{0, 1\}$. The range for MAP at $\theta = 0$ was 93% to 99%, and for ML the range was 81% to 94%; at $\theta = 1$, the range for MAP was 92% to 98%, and for ML it was 82% to 94%.

Variability measures for the ML/Alt and MAP/Alt procedures were comparable to their respective conventional procedures for tests of 25 items in length, a result also seen for the 15-item tests. The largest observed difference was 3% at $\theta = -2$ for the ML/Alt procedure, where the IQR was 2% as opposed to 5% for ML. Remaining differences between the alternative and conventional IQRs were negligible. The pattern of IQRs for GSS again closely mirrored that of ML/Alt, with nearly identical IQRs for $\theta > -1$, and slightly higher IQRs at $\theta = \{-2, -1\}$.

Maximum FII item selection

The efficiency measures from maximum FII item selection are summarized in Tables 9 and 10. Table 9 provides the medians and IQRs of the efficiency measures, and Table 10 provides the 25th and 75th percentile points of the efficiency measures. Table 11 provides the accumulated test information measures $I_0^{(T)}(\theta)$ for an optimal subset of items; this supplementary information may be used for converting the efficiency measures to accumulated test information measures.

Table 12 shows the difference in the median efficiency measures and the efficiency IQRs between maximum FI and maximum FII item selection. That is, Table 12 gives the arithmetic difference of Table 6 (maximum FI selection) from Table 9 (maximum FII selection). Those differences greater than 5% or less than −5% are considered relevant and are highlighted in the table. Notice that for tests of 15 and 25 items in length, item selection by maximum FII yielded similar (and in many cases nearly identical) median efficiency measures and efficiency IQRs as item selection by maximum FI. Because the efficiency measures under maximum FII and maximum FI item selection were so similar for these test lengths, discussion of the results will be limited to the shorter test lengths of 5 and 10 items.

Table 9. Medians and IQRs 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

Table 10. Efficiency measures at the 25th and 75th percentile points under maximum FII item selection.

		Ability level									
		θ = -2		θ = -1		θ = 0		θ = 1		θ = 2	
Ability estimation	Test length	P_{25}	P_{75}	P_{25}	P_{75}	P_{25}	P_{75}	P_{25}	P_{75}	P_{25}	P_{75}
ML	5	36.8	86.6	72.5	83.3	35.4	64.8	34.1	80.4	65.0	68.4
	10	76.1	94.1	74.9	89.7	50.6	78.5	56.0	82.7	73.8	87.7
	15	89.0	97.8	78.5	93.4	67.7	87.2	69.7	89.4	88.8	94.6
	25	93.4	98.9	87.3	96.4	82.4	93.5	83.2	93.6	94.5	97.4
ML/Alt	5	79.7	100.0	85.6	98.8	56.1	73.2	63.2	84.7	84.8	100.0
	10	87.3	99.6	81.0	97.0	62.4	86.4	68.4	92.7	83.7	100.0
	15	94.9	99.9	83.8	97.8	72.8	93.0	80.7	95.6	93.6	100.5
	25	93.8	99.0	89.9	98.6	85.7	96.4	89.2	98.8	97.2	99.8
MAP	5	8.1	57.3	78.9	98.8	77.2	92.9	77.9	95.9	32.1	32.1
	10	52.4	81.7	77.5	97.1	80.9	93.7	84.3	95.5	63.2	73.1
	15	72.6	91.4	82.4	97.8	87.2	98.2	87.6	97.5	82.2	89.2
	25	85.8	94.9	89.7	98.3	92.8	98.7	92.4	99.0	91.4	94.6
MAP/Alt	5	57.3	79.7	78.9	91.8	71.6	92.9	69.8	92.4	57.7	57.7
	10	62.9	90.6	77.2	96.1	78.9	93.7	83.0	94.5	64.3	82.8
	15	73.3	95.5	83.4	97.8	86.0	97.7	86.6	97.4	82.2	91.9
	25	85.3	97.4	90.2	98.3	92.2	98.5	91.6	98.9	91.4	95.7
GSS	5	79.7	96.7	68.2	89.5	63.3	83.1	63.2	84.7	71.6	98.4
	10	83.9	99.6	79.6	94.0	61.4	88.3	67.1	93.2	81.0	96.2
	15	92.7	98.0	83.9	95.8	70.0	92.3	77.4	93.8	92.3	98.7
	25	93.5	98.9	88.9	97.7	84.0	95.8	88.1	97.8	96.3	98.8

Table 11. Accumulated test information measures $I_0^{(T)}(\theta)$ for an optimal subset of items under maximum FII item selection.

	$I_0^{(T)}(\theta)$				
Test length	θ=-2	θ=-1	θ=0	θ=1	θ=2
5	2.351	2.768	5.282	6.333	6.043
10	4.697	5.478	9.571	12.836	12.237
15	6.457	7.862	13.039	18.558	16.398
25	9.351	11.843	18.569	28.12	23.527

Table 12. Differences in medians and IQRs of the efficiency measure between maximum FII and maximum FI item selection.

Ability estimation	Test length	Difference in median efficiency, Max FII − Max FI					Difference in efficiency IQR, Max FII − Max FI				
		$\theta=-2$	$\theta=-1$	$\theta=0$	$\theta=1$	$\theta=2$	$\theta=-2$	$\theta=-1$	$\theta=0$	$\theta=1$	$\theta=2$
ML	5	**29.8**	1.3	0	0.9	3.8	**5.6**	-1.3	0	0.9	**-13.7**
	10	**7.9**	1.8	0.2	2.3	2.6	-0.9	-4	-1	**-8.4**	0.6
	15	3.2	1.1	0.8	0.1	1.4	-0.3	-1	-2.3	-2.2	-1.3
	25	1.3	1.2	-0.3	1.1	0.7	0	-0.4	-1.2	-1.2	-0.3
ML/Alt	5	0	-4.4	**8.3**	-0.3	**-9.1**	1.7	**5.4**	**-9.6**	**-18.5**	-1.2
	10	-0.2	0.5	0.5	1.1	-3.1	0.6	-0.5	-3	**-6.3**	2.5
	15	0	-0.5	-0.2	1.4	-0.8	1	-0.3	-1.1	**-5.7**	1.9
	25	0	0.1	-0.6	0.6	-1	3	0.4	-1.9	-1.6	0.4
MAP	5	**11.8**	1.6	2.3	-3.5	**8.5**	0	0.3	0	**-5.7**	0
	10	**5.9**	-3.1	0.3	-1.3	**5.9**	0.2	-0.6	-2.1	-2.4	-3.7
	15	2.7	-0.8	0.3	0.1	0.5	0.5	-0.7	-1.3	-1.1	-2.1
	25	0.4	-0.3	-0.4	0.3	-0.1	1	-0.6	0	-0.9	0.1
MAP/Alt	5	0	-0.7	2.3	1.8	3.1	0	-2	0	0.4	0
	10	0.1	-1.7	0	-0.6	-3.2	**5.9**	-0.3	0.3	**-5.6**	-1.1
	15	0	-0.6	0.2	0.3	-0.6	3	-0.3	-0.5	-3.1	-0.5
	25	0	-0.3	-0.2	0.8	-0.5	3.4	0.3	-0.3	-1	1.1
GSS	5	0	1.5	-0.4	1.6	**-9.5**	0	0.3	0	**-10.2**	**23.9**
	10	3.4	-2.2	0	4.1	2.3	0	2.7	1	**-5**	2.9
	15	0	0.5	1	1.1	0.6	-1.9	-0.3	2.1	-1.9	0.1
	25	0	-0.3	0.2	0.5	0.3	-0.1	-0.9	-0.3	-0.4	-0.4

For each test length of 5 and 10 items, the following discussion begins with a comparison of the ability estimation procedures within maximum FII item selection, then makes comparisons between maximum FI and maximum FII item selection.

Test length of 5 items. Median efficiency measures and IQRs of the efficiency measures for tests of 5 items in length under maximum FII item selection are displayed in Figures 11 and 12. MAP was superior to ML for $\theta = \{-1, 0, 1\}$; however, ML was more efficient than MAP at the extreme ability levels $\theta = \{-2, 2\}$. MAP achieved a median efficiency exceeding 90% for $\theta = \{-1, 0, 1\}$, but performed poorly at the extremes, with median efficiencies of 43% and 32% at $\theta = \{-2, 2\}$, respectively. ML was less efficient than MAP at the middle of the ability continuum, having achieved a maximum median efficiency of 75% at $\theta = -1$ and minimum median efficiency of 46% at $\theta = 1$. Median efficiencies for ML at $\theta = \{-2, 2\}$ were 82% and 65%, respectively, and were higher than those observed for MAP.

The alternative ability estimation procedures performed better overall than the conventional ability estimation procedures for 5-item tests under maximum FII item selection. Median efficiencies for ML/Alt were greater than the corresponding measures observed for conventional ML across all ability levels. ML/Alt achieved a maximum median efficiency of 100% at $\theta = -2$ and a minimum median efficiency of 72% at $\theta = 0$. For MAP/Alt, the median efficiency measures were comparable to those observed for conventional MAP at $\theta = \{-1, 0, 1\}$, but median efficiency at the extremes was higher, at 80% and 58% for $\theta = \{-2, 2\}$, respectively. GSS performed similarly to ML/Alt for $\theta < 2$; however, at $\theta = 2$ the median efficiency of GSS was lower at 72%, as opposed to 85% for ML/Alt.

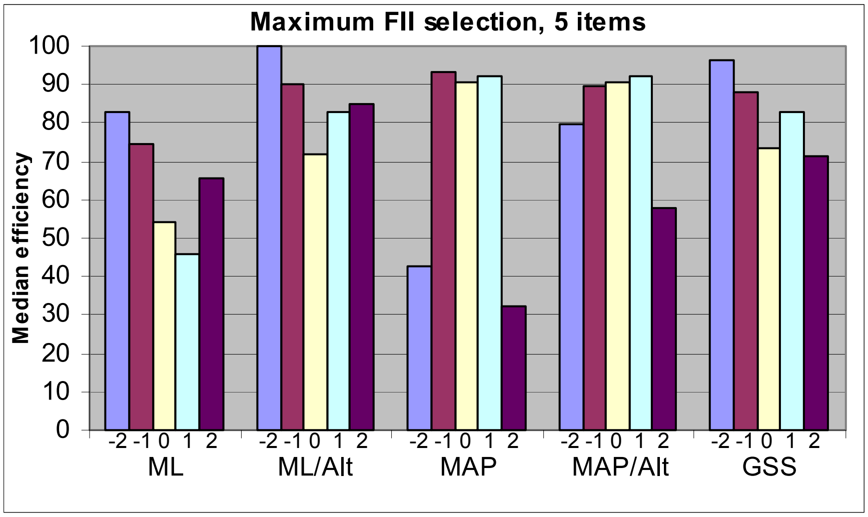

Figure 11. Median efficiency measures under maximum FII item selection for a test of 5 items.

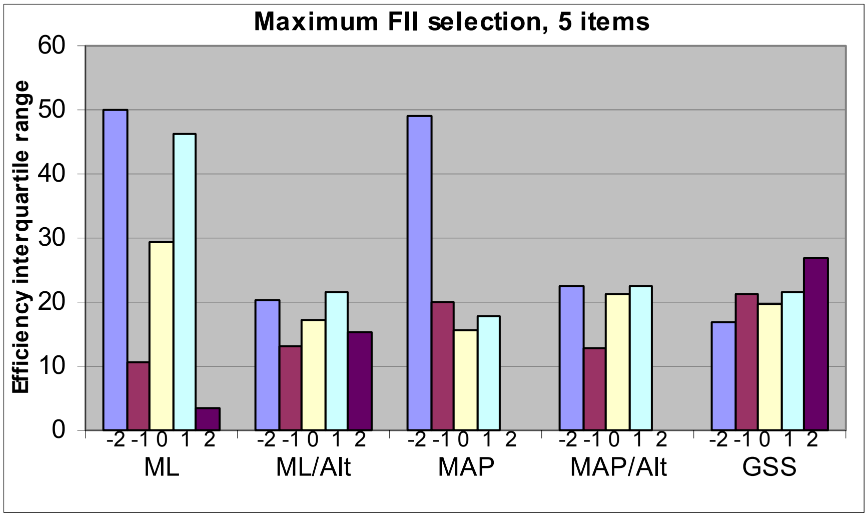

Figure 12. IQRs of the efficiency measure under maximum FII item selection for a test of 5 items. (Note: At θ = 2, the IQRs for MAP and MAP/Alt are equal to zero.)

For the conventional ability estimation procedures, the IQRs were smaller overall for MAP than ML, except at $\theta = \{-2, -1\}$. For MAP and $\theta > -1$, efficiency was least variable at $\theta = 2$ and most variable at $\theta = 1$; the respective IQRs were 0% and 15%. For ML and $\theta > -1$, efficiency was also least variable at $\theta = 2$ and most variable at $\theta = 1$; the respective IQRs were 3% and 46%. At $\theta = -1$, the IQR for MAP was higher than the IQR for ML, at 20% versus 11%. At the lowest extreme of ability, $\theta = -2$, the IQRs of efficiency for ML and MAP were 50% and 49%, respectively. If the 25th and 75th percentile points are considered at this ability level, then ML efficiency ranged from 37% to 87%, while MAP efficiency ranged from 8% to 57%. At the other extreme of ability, $\theta = 2$, ML also outperformed MAP, even though the IQR for MAP was 0%. The range in ML efficiency at $\theta = 2$ was 65% to 68%, while it was constant at 32% for MAP. For the middle range of ability, $-2 < \theta < 2$, MAP outperformed ML.

Variability in the efficiency measures from the alternative procedures was either comparable to or less than that observed for the conventional procedures. Reduction in variability was rather dramatic for ML/Alt and MAP/Alt at $\theta = -2$; for ML/Alt, the IQR was 20% in contrast to 50% for ML. Likewise, for MAP/Alt the IQR was 23%, in contrast to 49% for MAP. At $\theta = 2$, the IQR for MAP/Alt was equal to zero, a result also observed for conventional MAP. Variability in GSS was similar to that observed for MAP/Alt at ability levels $\theta = \{-2, 0, 1\}$. However, at $\theta = -1$ the IQR was higher at 22% as opposed to 13% for MAP. GSS was much more variable at $\theta = 2$, with an IQR of 27%.

For certain combinations of ability estimation procedure with true ability level, the efficiency measures under maximum FII were different than the corresponding measures under maximum FI selection. Examining first the conventional procedures, both ML and MAP benefited from maximum FII over maximum FI at the extreme ability levels $\theta = \{-2, 2\}$. Under maximum FI, the median efficiency measures at $\theta = -2$ for ML and MAP were 53% and 31%, respectively. Under maximum FII, the corresponding measures were 83% and 43%, respectively. Although the median efficiency measures were higher under FII selection, the measures at the 25th and 75th percentile points were approximately equal. For ML estimation under FII at $\theta = -2$, the measures at these percentile points were 37% and 87%, respectively; under FI, the measures were 37% and 81%, respectively. For MAP estimation under both FII and FI at $\theta = -2$, the range in efficiency was from 8% to 57%.

At the other extreme $\theta = 2$, under maximum FI the median efficiency measure for MAP was 23%, whereas under maximum FII it was 32%. For ML, the median efficiency measure was nearly equivalent under FI and FII, at 62% and 65%, respectively. Efficiency at the 25th and 75th percentile points for ML estimation under FII were higher at 65% and 68%, as opposed to 48% to 65% under FI. For MAP estimation under both FII and FI, the IQR was equal to zero.

The alternative estimation procedures did not uniformly benefit from maximum FII selection, however. While the median efficiency of ML/Alt was 9% higher (72% vs. 63%) at $\theta = 0$ under FII than under FI, it was 9% lower (85% vs. 94%) at $\theta = 2$. GSS also suffered at $\theta = 2$ under FII, with a median efficiency 9% lower (72% vs. 81%) than that under FI. Efficiency measures from the MAP/Alt procedure remained essentially unchanged under FII or FI selection.

The variability in efficiency for ML/Alt was reduced under FII at $\theta = \{0, 1\}$. Efficiency measures at the 25th and 75th percentile points for ML/Alt under FII at $\theta = 0$ were 56% and 73%, respectively; under FI, they were 45% and 72%, respectively. At $\theta = 1$, the corresponding efficiency measures for ML/Alt under FII were 63% and 85%, respectively; under FI, they were 47% and 87%, respectively. Variability was also reduced for GSS at $\theta = 1$ under FII, with the efficiency measures at the 25th and 75th percentile points being 63% and 85%, respectively; under FI, they were 51% and 83%, respectively. However, GSS not only suffered a reduction in median efficiency at $\theta = 2$ under FII, but also an increase in variability. Under maximum FI selection at this ability level, the measures were 78% and 81% at the two percentile points; under maximum FII selection, they were 71% and 98%. Oddly, the efficiency measure for GSS at $\theta = 2$ under FII was right-skewed. (The 25th and 50th percentile points were equal in this case.)

Test length of 10 items. Median efficiency measures and IQRs of the efficiency measures for tests of 10 items in length under maximum FII item selection are displayed in Figures 13 and 14. As observed for tests of 5 items in length, the conventional ability estimation procedures (ML and MAP) performed differently across the range of ability levels in terms of median efficiency. Again, MAP was superior to ML for $\theta = \{-1, 0, 1\}$ and ML was more efficient than MAP at the extreme ability levels $\theta = \{-2, 2\}$. MAP achieved a median efficiency at or above 91% for $\theta = \{-1, 0, 1\}$, but continued to lag in performance at the extremes, with median efficiencies of 79% and 70% at $\theta = \{-2, 2\}$, respectively.

The gap in median performance for MAP between the middle ability levels and those at the extremes narrowed from approximately 60% for 5-item tests to approximately 22% for 10-item tests. ML remained less efficient than MAP at the middle of the ability continuum, having achieved a maximum median efficiency of 86% at θ = −1 and minimum median efficiency of 72% at θ = 0. The median efficiency measures for ML at θ = {−2, 2} were again higher than those for MAP, at 90% and 83%, respectively.

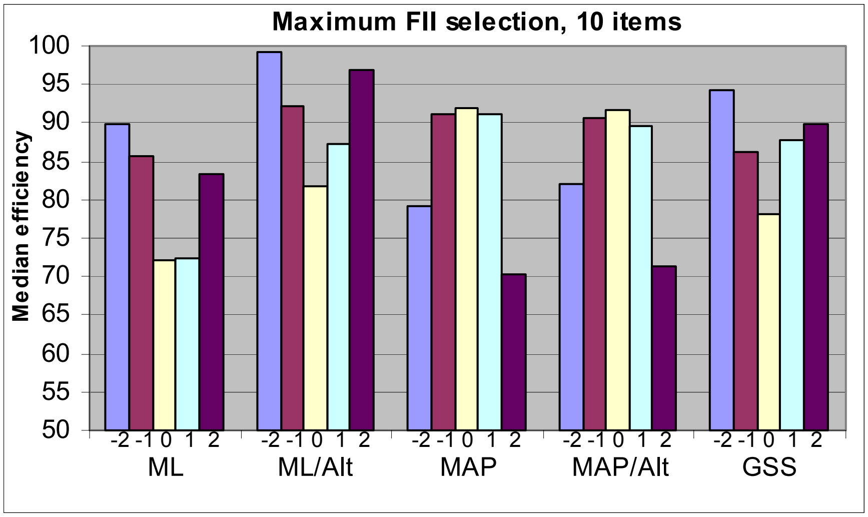

Figure 13. Median efficiency measures under maximum FII item selection for a test of 10 items.

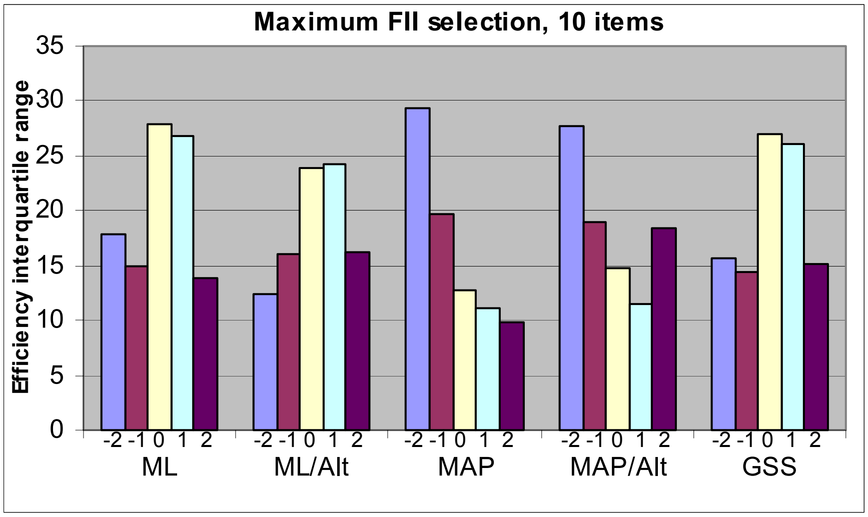

Figure 14. IQRs of the efficiency measure under maximum FII item selection for a test of 10 items.

Like the 5-item tests, the alternative ability estimation procedures performed better overall than the conventional ability estimation procedures for 10-item tests. Median efficiencies for ML/Alt were greater than the corresponding measures observed for conventional ML, with gains of approximately 10% across all ability levels. ML/Alt achieved a maximum median efficiency of 99% at $\theta = -2$, and a minimum median efficiency of 82% at $\theta = 0$. For MAP/Alt, the median efficiency measures were comparable to those observed for conventional MAP across all ability levels. GSS performed similarly to ML/Alt at $\theta = \{-2, 0, 1\}$, but with approximately 6% lower median efficiency measures than ML/Alt at $\theta = \{-1, 2\}$.

Unlike the 5-item tests, variability in the efficiency measures for the conventional ability estimation procedures for tests 10 items in length was substantially less for MAP than ML only at $\theta = 0$ and $\theta = 1$. At ability levels $\theta = \{0, 1\}$, the IQRs for MAP were 13% and 11%, respectively, whereas for ML they were 28% and 27%, respectively. At ability level $\theta = -1$, MAP was more variable than ML, with IQRs of 20% and 15%, respectively. At the extreme ability levels $\theta = \{-2, 2\}$, MAP was more variable than ML at $\theta = -2$, and less variable than ML at $\theta = 2$. IQRs at $\theta = -2$ for MAP and ML were 29% and 18%, respectively; at $\theta = 2$, they were 10% and 14%, respectively.

If the 25th and 75th percentile points are considered at $\theta = -2$, then ML efficiency ranged from 76% to 94%, while MAP efficiency ranged from 52% to 82%. At the other extreme of ability, $\theta = 2$, ML also outperformed MAP. The range in ML efficiency at $\theta = 2$ was 74% to 88%, while it was 63% to 73% for MAP. For the middle range of ability, $-2 < \theta < 2$, MAP outperformed ML.

Variability measures for the alternative procedures were somewhat different in the case of 10-item tests as compared to 5-item tests. Whereas IQRs were much lower at $\theta = -2$ for the alternative procedures than for the conventional procedures on 5-item tests, the reduction in variability was only about 6% for 10-item tests and only for ML/Alt. At $\theta = -2$, the IQR for ML/Alt was 12%, in contrast to 18% for ML. For ability levels $\theta > 0$, IQRs for ML/Alt were slightly smaller than those for ML. For MAP/Alt, variability was essentially unchanged at $\theta = \{-1, 0, 1\}$; at $\theta = 2$, the IQR for MAP/Alt was actually 9% higher. The pattern of IQRs for GSS closely mirrored that of ML/Alt, with nearly identical IQRs for $\theta > -1$, a slightly higher IQR at $\theta = -2$, and a slightly lower IQR at $\theta = -1$.

As observed for the 5-item tests, for certain combinations of ability estimation procedure with true ability level, the efficiency measures under maximum FII were different than the corresponding measures under maximum FI selection for tests of 10 items in length. For the conventional procedures, ML benefited from maximum FII over maximum FI item selection at $\theta = \{-2, 1\}$ and MAP benefited at both extreme ability levels $\theta = \{-2, 2\}$. Under maximum FI, the median efficiency measures at $\theta = -2$ for ML and MAP were 82% and 73%, respectively. Under maximum FII, the corresponding measures were 90% and 79%, respectively. Although the median efficiency measures were higher under FII selection, the measures at the 25th and 75th percentile points were approximately equal. For ML estimation under FII at $\theta = -2$, the measures at these percentile points were 76% and 94%, respectively; under FI, the measures were 73% and 91%, respectively. For MAP estimation under both FII and FI at $\theta = -2$, the range in efficiency was from 52% to 82%.

At the other extreme $\theta = 2$, under maximum FI the median efficiency measure for MAP was 64%, whereas under maximum FII it was 70%. For MAP under maximum FI at $\theta = 2$, the efficiency measures at the 25th and 75th percentile points were 59% and 73%, respectively; under FII, they were 63% and 73%, respectively. While the median efficiency of ML at $\theta = 1$ was similar under FI and FII, the variability was less under FII, with an IQR of 35% under FI and 27% under FII.

All three alternative ability estimation procedures appeared to benefit from maximum FII item selection in terms of reduced variability at ability level $\theta = 1$, although median efficiency measures remained essentially unchanged. The efficiency measures at the 25th to the 75th percentile points for ML/Alt at $\theta = 1$ under maximum FI were 64% and 94%, respectively; under maximum FII they were 68% to 93%, respectively. Likewise, for MAP/Alt under maximum FI, the corresponding measures were 77% and 95%, respectively; under maximum FII, they were 83% to 95%, respectively. For GSS under maximum FI, the corresponding measures were 62% and 93%, respectively; under maximum FII, they were 67% and 93%, respectively. However, at ability level $\theta = -2$, variability in the efficiency measures for MAP/Alt actually increased under maximum FII item selection. In this case, efficiency measures at the 25th and 75th percentile points under maximum FI were 68% and 91%, respectively, while under maximum FII they were 63% and 91%, respectively.

Test lengths of 15 and 25 items. As discussed earlier, the efficiency measures at these tests lengths under maximum FI and maximum FII item selection were very similar. Thus, the discussion of results under maximum FI selection suffices for maximum FII selection for 15- and 25-item tests. However, figures for the median efficiency and efficiency IQRs are provided for these test lengths. Figures 15 and 16 display these measures for 15-item tests; Figures 17 and 18 display these measures for 25-item tests.

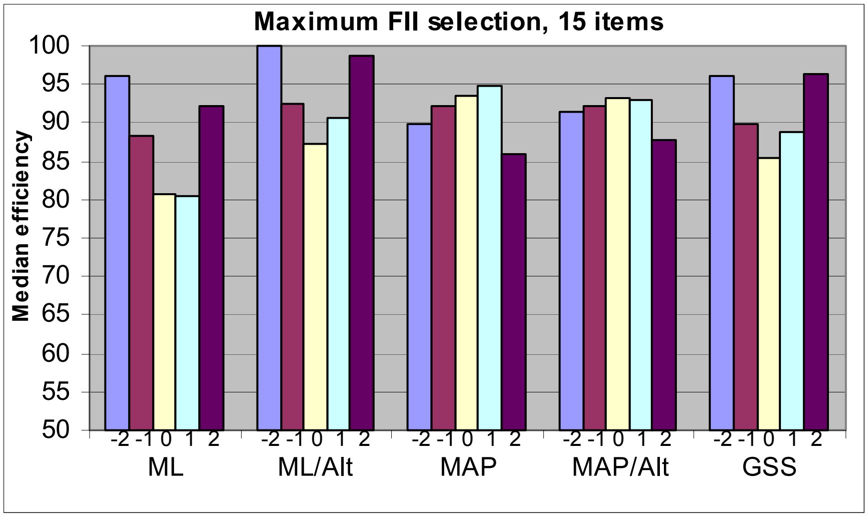

Figure 15. Median efficiency measures under maximum FII item selection for a test of 15 items.

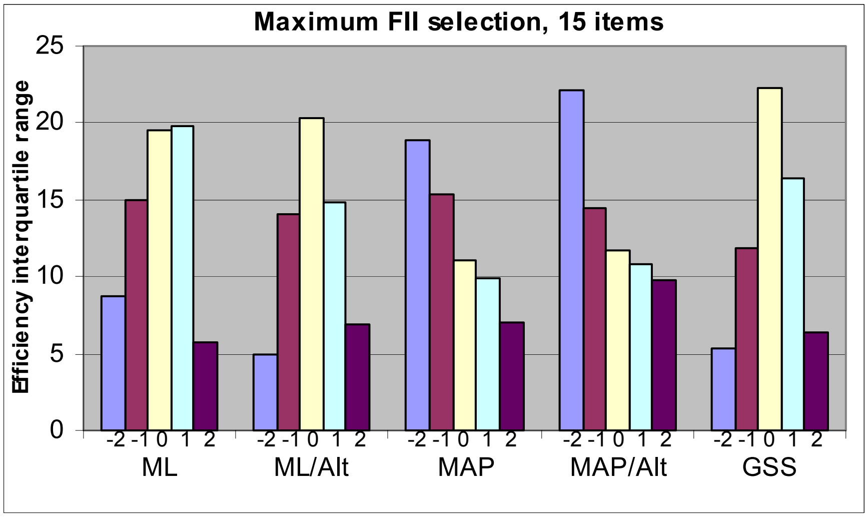

Figure 16. IQRs of the efficiency measure under maximum FII item selection for a test of 15 items.

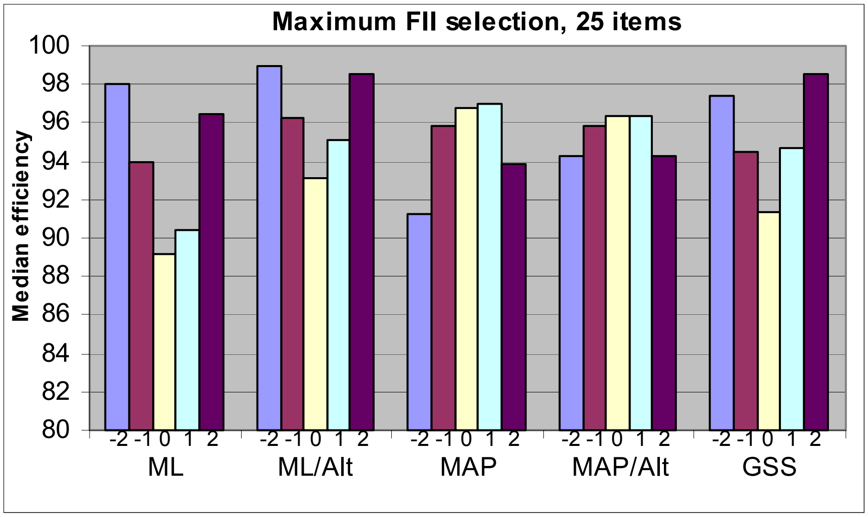

Figure 17. Median efficiency measures under maximum FII item selection for a test of 25 items.

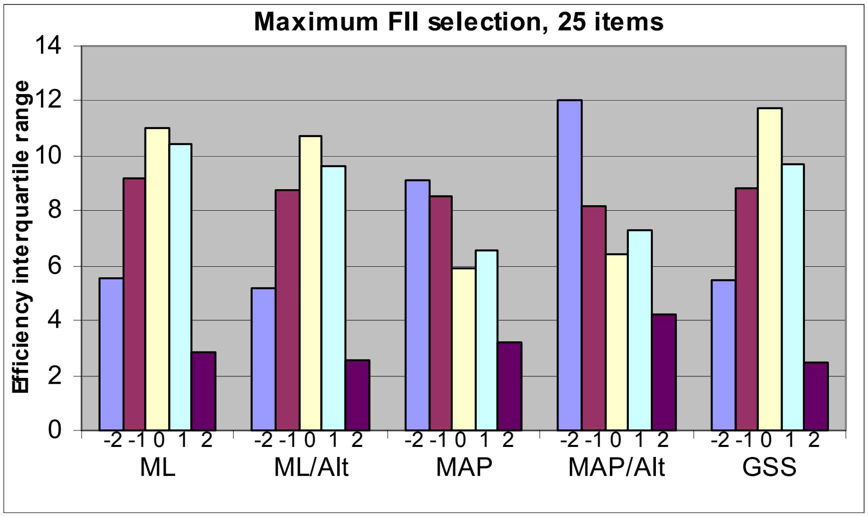

Figure 18. IQRs of the efficiency measure under maximum FII item selection for a test of 25 items.

Provisional ability estimates

Summary information for provisional ability estimates at each level of ability estimation procedure × test length × true ability level is provided in Tables 13 and 14 for item selection under maximum FI, and in Tables 15 and 16 for item selection under maximum FII. Means and medians of the provisional ability estimates are provided in Tables 13 and 15; mean-squared errors and standard deviations are provided in Tables 14 and 16.

Table 13. Means and medians of provisional ability estimates under maximum FI item selection.

Ability estimation	Test length	Mean $\hat{\theta}$					Median $\hat{\theta}$				
		θ=-2	θ=-1	θ=0	θ=1	θ=2	θ=-2	θ=-1	θ=0	θ=1	θ=2
ML	5	-2.060	-1.025	0.142	1.130	2.098	-2.047	-1.214	0.115	1.054	2.014
	10	-2.012	-0.988	0.058	1.059	2.040	-2.002	-1.003	0.077	1.048	1.985
	15	-2.002	-0.989	0.032	1.034	2.022	-2.002	-0.994	0.030	1.024	2.006
	25	-2.010	-1.003	0.010	1.015	2.004	-2.009	-1.005	0.011	1.012	1.997
ML/Alt	5	-1.607	-0.927	0.105	1.008	1.900	-1.744	-1.044	0.050	1.063	2.085
	10	-1.683	-0.901	0.133	1.009	1.902	-1.794	-0.972	0.079	1.007	2.004
	15	-1.636	-0.881	0.144	1.014	1.917	-1.814	-0.975	0.083	1.023	1.978
	25	-1.588	-0.869	0.148	1.015	1.855	-1.728	-0.957	0.100	1.003	1.900
MAP	5	-1.202	-0.745	0.036	0.882	1.509	-1.161	-0.634	0.029	0.875	1.580
	10	-1.543	-0.837	0.027	0.936	1.821	-1.587	-0.876	0.036	0.935	1.769
	15	-1.682	-0.883	0.015	0.955	1.889	-1.694	-0.885	0.008	0.964	1.881
	25	-1.791	-0.928	0.006	0.967	1.922	-1.803	-0.917	-0.004	0.970	1.913
MAP/Alt	5	-1.548	-0.846	0.028	0.911	1.665	-1.738	-0.705	0.029	0.849	1.429
	10	-1.638	-0.857	0.026	0.944	1.807	-1.781	-0.851	0.043	0.935	1.753
	15	-1.633	-0.876	0.021	0.959	1.889	-1.809	-0.886	0.016	0.964	1.912
	25	-1.694	-0.917	0.014	0.972	1.884	-1.778	-0.923	0.009	0.975	1.927
GSS	5	-2.015	-1.033	0.058	1.015	2.086	-2.472	-1.167	0.000	0.807	2.249
	10	-1.961	-1.026	0.010	1.008	2.000	-1.889	-1.167	0.000	1.029	1.889
	15	-1.969	-1.010	0.018	1.004	2.024	-1.889	-1.167	0.000	1.167	1.889
	25	-1.988	-1.002	0.003	1.001	2.006	-1.889	-1.167	0.000	1.029	1.889

Table 14. Mean-squared errors and standard deviations of provisional ability estimates under maximum FI item selection.

Ability estimation	Test length	MSE $\left(\hat{\theta}-\theta\right)^2$					S.D. $\hat{\theta}$				
		θ=-2	θ=-1	θ=0	θ=1	θ=2	θ=-2	θ=-1	θ=0	θ=1	θ=2
ML	5	1.010	0.713	0.615	0.419	0.398	1.003	0.844	0.771	0.634	0.623
	10	0.456	0.293	0.206	0.146	0.177	0.675	0.541	0.450	0.378	0.419
	15	0.258	0.172	0.119	0.080	0.090	0.508	0.415	0.343	0.281	0.300
	25	0.139	0.095	0.067	0.046	0.048	0.373	0.309	0.258	0.214	0.218
ML/Alt	5	0.436	0.441	0.404	0.285	0.137	0.531	0.660	0.627	0.534	0.356
	10	0.310	0.317	0.212	0.125	0.080	0.458	0.554	0.441	0.353	0.265
	15	0.292	0.275	0.164	0.074	0.054	0.400	0.511	0.378	0.271	0.217
	25	0.302	0.241	0.130	0.042	0.041	0.364	0.473	0.329	0.205	0.140
MAP	5	0.847	0.251	0.172	0.156	0.281	0.458	0.431	0.413	0.377	0.200
	10	0.395	0.170	0.108	0.087	0.112	0.431	0.379	0.328	0.288	0.282
	15	0.249	0.122	0.079	0.064	0.075	0.384	0.329	0.281	0.248	0.251
	25	0.133	0.077	0.053	0.041	0.046	0.299	0.268	0.230	0.200	0.201
MAP/Alt	5	0.572	0.358	0.189	0.184	0.250	0.606	0.578	0.434	0.420	0.371
	10	0.350	0.193	0.112	0.092	0.105	0.468	0.415	0.334	0.298	0.260
	15	0.262	0.137	0.083	0.065	0.062	0.357	0.349	0.288	0.251	0.222
	25	0.166	0.095	0.056	0.042	0.035	0.269	0.297	0.237	0.203	0.146
GSS	5	0.762	0.517	0.446	0.285	0.306	0.873	0.718	0.665	0.534	0.546
	10	0.348	0.252	0.206	0.133	0.144	0.589	0.501	0.454	0.364	0.379
	15	0.242	0.182	0.134	0.086	0.097	0.491	0.426	0.365	0.293	0.311
	25	0.165	0.112	0.083	0.052	0.055	0.406	0.335	0.288	0.229	0.235

Table 15. Means and medians of provisional ability estimates under maximum FII item selection.

Ability estimation	Test length	Mean $\hat{\theta}$					Median $\hat{\theta}$				
		θ=-2	θ=-1	θ=0	θ=1	θ=2	θ=-2	θ=-1	θ=0	θ=1	θ=2
ML	5	-2.017	-0.987	0.151	1.117	2.095	-1.912	-1.214	0.115	1.054	2.014
	10	-2.006	-0.980	0.050	1.049	2.037	-2.002	-0.984	0.046	1.036	1.961
	15	-2.000	-0.992	0.031	1.032	2.023	-1.994	-0.993	0.027	1.029	1.997
	25	-2.009	-1.006	0.013	1.012	2.004	-2.006	-1.005	0.015	1.012	1.997
ML/Alt	5	-1.600	-0.921	0.091	1.004	1.897	-1.729	-1.044	0.105	0.908	1.932
	10	-1.653	-0.899	0.115	1.01	1.905	-1.771	-0.975	0.083	0.993	1.957
	15	-1.625	-0.878	0.128	1.015	1.918	-1.760	-0.963	0.091	1.016	1.967
	25	-1.572	-0.862	0.148	1.011	1.837	-1.728	-0.954	0.110	1.005	1.862
MAP	5	-1.202	-0.743	0.035	0.872	1.553	-1.167	-0.832	0.035	0.875	1.655
	10	-1.550	-0.838	0.025	0.939	1.826	-1.572	-0.857	0.021	0.928	1.827
	15	-1.683	-0.885	0.016	0.950	1.888	-1.691	-0.899	0.013	0.956	1.883
	25	-1.789	-0.928	0.009	0.965	1.924	-1.797	-0.920	0.003	0.965	1.915
MAP/Alt	5	-1.523	-0.838	0.033	0.915	1.666	-1.738	-0.705	0.035	0.849	1.429
	10	-1.619	-0.855	0.022	0.941	1.805	-1.883	-0.824	0.022	0.935	1.791
	15	-1.616	-0.876	0.023	0.954	1.884	-1.809	-0.894	0.019	0.956	1.918
	25	-1.678	-0.917	0.018	0.969	1.871	-1.778	-0.929	0.014	0.970	1.927
GSS	5	-2.018	-1.039	0.060	1.026	2.079	-2.472	-1.167	0.000	0.807	1.889
	10	-1.943	-1.028	0.024	1.008	2.004	-1.889	-1.167	0.000	1.167	1.889
	15	-1.972	-1.015	0.017	1.007	2.021	-1.889	-1.167	0.000	1.029	1.889
	25	-1.986	-1.005	0.010	1.007	2.002	-1.889	-1.167	0.000	1.029	1.889

Table 16. Mean-squared errors and standard deviations of provisional ability estimates under maximum FII item selection.

Ability estimation	Test length	MSE $\left(\hat{\theta}-\theta\right)^2$					S.D. $\hat{\theta}$				
		θ=-2	θ=-1	θ=0	θ=1	θ=2	θ=-2	θ=-1	θ=0	θ=1	θ=2
ML	5	0.924	0.645	0.591	0.398	0.390	0.961	0.803	0.754	0.620	0.617
	10	0.437	0.257	0.204	0.135	0.151	0.661	0.507	0.449	0.364	0.387
	15	0.252	0.162	0.120	0.079	0.087	0.502	0.402	0.345	0.280	0.294
	25	0.133	0.097	0.066	0.046	0.046	0.365	0.311	0.257	0.213	0.215
ML/Alt	5	0.437	0.426	0.372	0.255	0.131	0.526	0.648	0.603	0.505	0.347
	10	0.320	0.299	0.196	0.112	0.070	0.447	0.537	0.427	0.334	0.246
	15	0.300	0.274	0.143	0.070	0.047	0.399	0.509	0.356	0.265	0.201
	25	0.318	0.251	0.121	0.042	0.040	0.367	0.482	0.314	0.204	0.118
MAP	5	0.854	0.248	0.161	0.147	0.253	0.466	0.427	0.400	0.361	0.231
	10	0.393	0.164	0.108	0.083	0.112	0.437	0.371	0.328	0.282	0.285
	15	0.246	0.121	0.079	0.061	0.073	0.381	0.329	0.280	0.242	0.245
	25	0.133	0.076	0.054	0.040	0.047	0.298	0.266	0.232	0.198	0.204
MAP/Alt	5	0.626	0.351	0.183	0.179	0.248	0.631	0.570	0.427	0.414	0.369
	10	0.384	0.187	0.113	0.089	0.105	0.489	0.407	0.336	0.292	0.259
	15	0.287	0.134	0.080	0.062	0.058	0.374	0.344	0.282	0.244	0.212
	25	0.183	0.096	0.056	0.041	0.035	0.282	0.298	0.236	0.200	0.136
GSS	5	0.745	0.506	0.442	0.292	0.249	0.863	0.710	0.662	0.540	0.493
	10	0.344	0.245	0.194	0.138	0.135	0.584	0.494	0.440	0.371	0.367
	15	0.237	0.174	0.140	0.086	0.091	0.486	0.417	0.374	0.293	0.301
	25	0.167	0.111	0.082	0.052	0.052	0.409	0.333	0.286	0.228	0.229

For a given level of ability estimation procedure × test length, mean provisional ability estimates under maximum FI and maximum FII item selection are comparable; the same holds true for the standard deviations of the provisional ability estimates. With respect to the mean provisional ability estimates, those from the ML and GSS ability estimation procedures appear to be the least biased for all test lengths. Estimates showing the largest amount of bias are MAP, followed by MAP/Alt and ML/Alt. The inward bias of the MAP estimates may be attributed to the *N*(0,1) prior. Because the distributions of provisional ability estimates from the ML/Alt and MAP/Alt estimation procedures are right-skewed, the means of these distributions should be interpreted with caution as measures of central tendency. Median measures for these distributions are provided to illustrate this effect; note that the bias is typically less pronounced for the medians of the distributions from ML/Alt and MAP/Alt.

The standard deviation of the provisional ability estimates is generally smallest for MAP estimates, as the presence of an informative prior increases the precision of these estimates. For tests 5 items in length, the ML provisional estimates are the most variable among all the ability estimation procedures, although the GSS procedure is only slightly less variable. ML/Alt estimates are considerably less variable than ML estimates at the extreme ability levels $\theta = \{-2, 2\}$. MAP/Alt estimates are generally more variable than the respective MAP estimates for test lengths of 5 and 10 items. Again, it should be noted that the distributions of ability estimates from the ML/Alt and MAP/Alt procedures are right-skewed, and so the standard deviation of these estimates may not accurately characterize the variability in estimates.

In addition to the standard deviation of the provisional ability estimates, the mean-squared errors of the provisional ability estimates are also provided. As observed for the standard deviation of these estimates, for a given level of ability estimation procedure × test length, the mean squared errors under maximum FI and maximum FII item selection are comparable. For the shorter test lengths (5 or 10 items), the mean-squared errors are smallest for MAP and MAP/Alt.

CHAPTER 5

Summary and Conclusions

Summary

Efficiency is often cited as an advantage of computerized adaptive tests (CATs) over traditional paper-and-pencil tests. The goal of a CAT is to administer items targeted to examinee ability, where higher-ability examinees generally receive more difficult items and lower-ability examinees generally receive less difficult items. Nevertheless, item selection in a CAT at the early stages of test administration has been criticized as being inefficient, as provisional ability estimates are typically imprecise, inaccurate, or both. The argument contends that because item selection is dependent on ability estimation, item selection based on these early provisional ability estimates is likely to be mismatched with respect to an examinee's true ability.

The efficiency of CAT item selection is dependent on item selection procedures as well as ability estimation procedures. Most commonly, maximum Fisher information (FI) item selection is employed in conjunction with either maximum likelihood (ML) or modal a posteriori (MAP) ability estimation. Because maximum FI item selection (under either ML or MAP) has been criticized as being inefficient, the first purpose of this study was to quantify the efficiency (or, inefficiency) of this most common item selection procedure. The second purpose of this study was to propose an alternative ability estimation procedure that addresses potential inefficiencies in CAT item selection, where this alternative procedure operates concurrently with either ML or MAP estimation and functions as an adjustment to either of these procedures. The third purpose of this study was to evaluate the efficiency of CAT item selection given five ability estimation procedures (i.e., ML, ML/Alt, MAP, MAP/Alt, and GSS, where ML/Alt uses the alternative procedure concurrently with ML estimation, MAP/Alt uses the alternative procedure concurrently with MAP estimation, and GSS is a golden section search strategy), with two item selection procedures (i.e., maximum FI and maximum Fisher interval information, or FII).

Further, this study utilized a precise definition for an efficiency measure. The primary advantage of this definition was that the efficiency of item selection from different procedures (e.g., alternative item selection procedures or alternative ability estimation procedures) could be compared to a fixed point of reference, one which characterizes the most efficient estimator possible.

Two primary research questions were thus investigated in this study:

1. How might the efficiency of maximum FI item selection under conventional ability estimation procedures be characterized, especially at the early stages of a CAT administration? More specific questions include: (a) After a fixed number of items have been administered, to what extent does efficiency of maximum FI item selection under ML or MAP ability estimation vary for different points along the ability continuum? (b) What is the effect of ability estimation procedure on the efficiency of maximum FI item selection?

2. Is it possible to improve upon the efficiency of maximum FI item selection under conventional ability estimation procedures by utilizing alternative item selection procedures, alternative ability estimation procedures, or a combination of both? More

specific questions include: (a) After a fixed number of items have been administered, to what extent do the efficiency measures for the alternative item selection and ability estimation procedures vary for different points along the ability continuum? Specifically, how do the alternatives to FI item selection with ML or MAP ability estimation compare to one another? (b) How do these efficiency measures compare with those obtained for maximum FI item selection with ML or MAP ability estimation? That is, to what extent are the alternative item selection and ability estimation procedures more (or less) efficient than maximum FI item selection in conjunction with conventional ability estimation procedures?

The two primary research questions were addressed using a simulation methodology. The CAT simulations employed here draw on an item bank of 367 pre-calibrated and dichotomously-scored 3P items from a recently-administered large-scale CAT assessment of mathematics ability. In its operational form, the CAT administered using this item bank is fixed at a length of 28 items; however, as it was hypothesized that the greatest variation in CAT efficiency would occur much earlier (e.g., at or before the 10th administered item), the CAT simulations were fixed such that no test exceeded a length of 25 items.

The four factors in the experimental design were: (1) item selection procedure (maximum FI or maximum FII item selection); (2) ability estimation procedure (ML, ML/Alt, MAP, MAP/Alt, and GSS); (3) true ability level at discrete points along the ability continuum (at -2, -1, 0, +1, or +2 logits); and (4) test length (5, 10, 15, or 25 items). For each of the experimental conditions, 1000 replications were generated. Efficiency was the primary dependent measure. Since this measure is highly skewed to the left, the median efficiency was reported as a measure of central tendency, and the interquartile range was reported as a measure of variability.

With respect to the first research question, sizeable differences in efficiency were found between ML and MAP ability estimation under maximum FI item selection for shorter tests (5 or 10 items) across true ability levels. At the middle of the ability distribution, MAP was more efficient; at the extremes of the ability distribution; ML was more efficient. For longer tests (15 or 25 items), these differences remained but became far less profound.

With respect to the second research question, increased test efficiency was obtained using alternative ability estimation procedures (ML/Alt, MAP/Alt, and GSS) in conjunction with maximum FI item selection. The gains in efficiency were most pronounced for shorter tests, but were noticeable even for longer tests. Conventional ability estimation procedures (ML and MAP) benefited from the alternative item selection procedure (maximum FII selection) at the extremes of the ability distribution for shorter tests, but efficiency measures as compared to maximum FI selection were essentially unchanged for longer tests. Mixed results occurred when maximum FII selection was combined with the alternative ability estimation procedures. However, as observed for the conventional ability estimation procedures under maximum FII selection, there was no change in efficiency for longer test lengths.

Findings and Conclusions

The efficiency measure $I_{CAT}^{(T)}(\theta)/I_0^{(T)}(\theta)$ used in this study plays an important role in the interpretation of the results. Comparisons across procedures certainly could have been made if relative efficiency measures $I_A^{(T)}(\theta)/I_B^{(T)}(\theta)$ for two procedures *A* and *B* were used instead. However, by utilizing $I_0^{(T)}(\theta)$, the upper bound on the precision with which an examinee with true ability θ may be measured, both relative and absolute comparisons become possible. Thus, not only can two procedures be characterized in terms of how much more efficient one is than the other, but also in terms of how efficient each one is with respect to the maximum efficiency attainable.

The first purpose of this study was to quantify the efficiency of the most common item selection procedure, maximum FI item selection, in conjunction with ML and MAP ability estimation procedures. From a relative efficiency standpoint, it was found that 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. However, the quantification of efficiency indicates how well, in terms of optimal performance, the procedures are operating. 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.

Maximum FII item selection was proposed to address the imprecision in ability estimation at the early stages of a CAT. Some of the results from maximum FII item selection in conjunction with ML and MAP estimation from this study 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. However, 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.

As mentioned, the room for improvement in efficiency exists for ML ability estimation, across all levels of ability, and for MAP at the extremes. In addition, the largest gaps in performance occurred for the shorter test lengths. Maximum FII item selection filled in some of these gaps, raising 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.

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 third ability estimation procedure, GSS, had been previously investigated by Xiao (1999). Xiao (1999) states that the ability estimate from the GSS procedure is equivalent to the ML estimate; however, a comparison of the results from the ML and GSS procedures in this study suggests that they are not. Were the estimates the same, the expectation would be that the efficiency measures from ML and GSS would also be the same, or at least very similar. However, the 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 not from ML, but from ML/Alt. This correspondence may result from the fact that GSS, like ML/Alt, utilizes hypothesis-testing and an interval search strategy.

When maximum FII item selection was combined with the alternative ability estimation procedures, 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.

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.

Implications for Further Study

There are two major areas where the present study could be extended: (1) investigations relating to the robustness of item selection under more realistic testing conditions; and (2) the formulation of the alternative ability estimation procedure. With respect to the first area, it should be noted that the present study investigated CAT item selection in a highly idealized case, where all item responses were simulated according to a unidimensional IRT model. However, actual test data is almost always multidimensional, containing more noise than is generated by unidimensional 3P simulations. An important extension to the study would be to simulate examinee response data under a more realistic model, such as a multidimensional IRT model, but administer a unidimensional CAT. Thus, while the items within the CAT pool would be unidimensional, examinee responses would contain extraneous sources of variance. The performance of the ability estimation and item selection procedures under these more realistic conditions could then be examined, and the more robust procedures identified.

In addition, the efficiency of item selection could be investigated in the more realistic situation where item exposure control is utilized. Item exposure control necessarily reduces test efficiency, because the "best" item selected by the CAT algorithm cannot always be administered. The efficiency measure can indicate to what extent efficiency is lowered from its optimal values (i.e., efficiency when no item exposure control is operating).

The second area for further research concerns the formulation of the alternative ability estimation procedure. The procedure itself contains two components: a hypothesis test and a search procedure. Currently, the hypothesis test compares expected and observed proportions correct, and the decision rule is based on a critical *z*-value z_c. The optimal z_c values were found empirically in this study; future research could investigate whether these z_c values remain constant or vary under different conditions, such as for different item pool sizes and different characteristics of the items in the pool. This research might also yield clues for an analytic solution to finding optimal z_c values. In addition, while the current hypothesis test compares expected and observed proportions correct, expected and observed likelihood functions could be compared instead, in a similar manner as Drasgow, Levine, & Williams' (1985) standardized likelihood-based statistic l_z.

The second component to the alternative estimation procedure is the search for a new ability estimate. The current procedure utilizes the average ICC for the set of items already administered to find this new ability estimate, but other possibilities exist. For example, the average ICC is essentially a test score where item probabilities are all weighted equally. Another scoring convention, where the weights are not all constant, might be applied. Such a technique was utilized in the GSS procedure, where optimal scoring weights were applied to examinees' response vectors. Further, if a test statistic similar to l_z is used for hypothesis-testing, then it may become possible to use likelihood functions to search for the new ability estimate, as opposed to average ICCs or related score functions.

APPENDIX A

Item parameters from CAT pool

Table 17. Item parameters from CAT pool

Item number	a	b	c
1	1.688	0.783	0.261
2	0.953	-0.016	0.125
3	0.675	-1.117	0.000
4	0.924	-0.440	0.128
5	0.872	-0.452	0.151
6	0.798	0.257	0.083
7	0.727	-1.415	0.000
8	0.926	-1.860	0.132
9	0.843	1.098	0.166
10	1.528	1.322	0.231
11	0.811	-1.047	0.174
12	0.749	0.383	0.065
13	0.907	1.424	0.076
14	0.855	-1.700	0.000
15	0.770	0.647	0.129
16	1.009	-0.106	0.342
17	0.611	-2.844	0.110
18	0.984	0.505	0.105
19	0.586	0.057	0.000
20	0.880	0.433	0.060
21	0.632	-0.891	0.000
22	1.144	-0.719	0.254
23	1.001	1.572	0.086
24	0.605	-2.273	0.123
25	1.246	0.850	0.244
26	0.944	1.938	0.127
27	1.449	1.144	0.167
28	1.106	0.199	0.227
29	0.810	-1.683	0.000
30	0.891	0.001	0.054
31	0.921	0.448	0.082
32	0.723	-1.903	0.101
33	0.646	0.571	0.073
34	1.290	-0.344	0.077
35	1.937	1.988	0.150
36	1.121	-0.446	0.196
37	0.588	-1.275	0.096
38	0.506	0.028	0.000
39	1.175	2.049	0.131
40	0.706	-1.027	0.070
41	1.156	1.178	0.339
42	1.599	1.243	0.235
43	0.700	-0.504	0.000
44	1.530	0.300	0.103
45	0.907	-0.205	0.052
46	1.124	1.922	0.148
47	1.341	0.342	0.110
48	1.595	1.039	0.142
49	0.834	-0.349	0.069
50	1.041	-0.124	0.209
51	1.302	-0.211	0.074
52	1.088	-0.048	0.328
53	1.194	0.829	0.018
54	0.673	-1.049	0.089
55	1.926	1.156	0.203
56	1.541	0.800	0.116
57	0.658	-2.103	0.036
58	1.947	0.497	0.239
59	1.812	1.469	0.122
60	0.864	1.183	0.112
61	1.935	0.872	0.247
62	0.702	-0.365	0.000
63	1.352	1.862	0.195
64	0.923	-2.054	0.134
65	0.831	-0.388	0.016
66	1.353	-0.329	0.241
67	1.508	2.202	0.059
68	0.667	-0.878	0.000
69	1.760	1.956	0.192
70	0.642	0.168	0.034
71	0.854	-0.520	0.000
72	1.926	1.341	0.282
73	1.265	1.200	0.159
74	1.206	1.012	0.098
75	1.274	-0.804	0.193
76	1.425	0.902	0.123
77	1.695	1.710	0.209
78	1.129	0.426	0.159
79	1.444	2.105	0.147
80	0.950	0.563	0.000
81	1.486	0.503	0.279
82	1.361	0.888	0.133
83	1.197	0.943	0.119
84	1.361	1.331	0.080
85	1.407	1.282	0.101
86	0.914	-0.113	0.128
87	0.979	0.353	0.079
88	1.003	-0.014	0.113
89	1.037	-0.644	0.187
90	0.969	-1.445	0.000
91	0.663	-1.288	0.095
92	0.889	0.344	0.103
93	0.730	0.158	0.000
94	1.930	0.926	0.316
95	1.233	0.355	0.104
96	0.961	0.311	0.102

Table 17 (cont'd).

Item number	a	b	c
97	1.648	0.963	0.223
98	1.282	0.905	0.106
99	1.775	0.617	0.108
100	1.387	-0.093	0.216
101	1.026	-0.259	0.225
102	0.886	-0.987	0.000
103	0.931	-0.721	0.067
104	0.645	-0.423	0.124
105	1.624	1.606	0.164
106	0.478	-2.054	0.095
107	0.977	0.026	0.156
108	0.465	-0.210	0.142
109	0.756	-0.912	0.181
110	0.476	-0.255	0.050
111	0.498	-0.980	0.072
112	1.124	0.540	0.108
113	1.229	1.607	0.047
114	1.509	1.692	0.075
115	1.050	0.209	0.141
116	0.951	2.054	0.098
117	1.090	1.850	0.229
118	0.697	-1.836	0.101
119	0.771	2.743	0.058
120	0.677	-1.219	0.000
121	0.700	0.104	0.000
122	0.600	-0.483	0.000
123	0.838	-0.212	0.123
124	0.868	0.286	0.026
125	0.654	-0.817	0.000
126	0.530	-1.529	0.123
127	0.842	0.282	0.102
128	0.909	-1.375	0.023
129	0.718	0.316	0.094
130	0.798	0.756	0.150
131	0.662	-1.200	0.031
132	0.951	-1.767	0.023
133	0.963	0.995	0.029
134	0.989	0.745	0.264
135	0.631	-0.013	0.077
136	0.591	-2.069	0.133
137	1.888	1.280	0.183
138	1.159	0.392	0.288
139	0.730	-0.753	0.056
140	1.105	0.004	0.115
141	0.740	-1.217	0.000
142	0.812	0.353	0.333
143	1.376	0.043	0.197
144	0.675	0.012	0.074
145	1.063	-0.160	0.404
146	1.011	1.676	0.269
147	0.569	-0.774	0.004
148	0.900	-0.628	0.335
149	1.529	1.211	0.175
150	0.654	-1.206	0.142
151	1.006	-0.404	0.197
152	0.787	-0.961	0.120
153	0.700	-0.684	0.069
154	0.874	-1.549	0.000
155	0.624	-0.538	0.249
156	1.139	-0.734	0.202
157	1.004	-1.031	0.305
158	1.008	1.376	0.172
159	1.004	-1.668	0.000
160	1.356	0.582	0.337
161	0.988	-1.208	0.347
162	1.110	1.122	0.118
163	0.669	-2.519	0.133
164	0.860	0.314	0.156
165	0.936	0.672	0.037
166	0.793	-0.817	0.198
167	1.213	-0.762	0.226
168	1.127	0.430	0.359
169	1.029	0.650	0.123
170	1.253	1.817	0.091
171	1.009	-0.193	0.253
172	1.410	0.579	0.318
173	0.806	-1.838	0.000
174	0.943	0.146	0.259
175	0.917	0.241	0.077
176	1.184	1.341	0.192
177	1.775	2.018	0.218
178	0.906	-0.639	0.000
179	0.937	0.053	0.216
180	0.784	-1.696	0.133
181	1.260	1.923	0.362
182	0.588	-0.783	0.088
183	1.168	0.927	0.171
184	1.114	0.564	0.234
185	0.652	0.150	0.041
186	0.767	-0.493	0.267
187	0.836	-0.305	0.042
188	0.912	-0.873	0.391
189	1.044	0.858	0.036
190	1.327	0.550	0.327
191	0.848	-0.213	0.083
192	1.083	1.052	0.288

Table 17 (cont'd).

Item number	a	b	c
193	1.584	0.731	0.172
194	0.633	-0.967	0.000
195	1.353	0.787	0.306
196	1.335	1.383	0.132
197	1.097	-0.410	0.436
198	1.440	0.580	0.330
199	1.390	0.228	0.473
200	0.504	-0.530	0.000
201	0.634	-1.293	0.120
202	0.709	-0.140	0.032
203	0.891	-1.080	0.336
204	0.991	0.850	0.034
205	0.578	-0.785	0.010
206	1.377	0.099	0.204
207	0.887	-1.221	0.231
208	1.092	1.422	0.063
209	1.396	-0.075	0.157
210	1.091	0.372	0.315
211	1.120	0.286	0.292
212	0.883	-0.451	0.294
213	0.563	-0.823	0.133
214	0.939	0.042	0.294
215	1.060	1.131	0.029
216	1.102	0.754	0.283
217	0.715	-0.078	0.000
218	0.572	-1.186	0.000
219	0.692	-0.957	0.000
220	0.868	-0.176	0.104
221	0.710	-1.470	0.001
222	0.843	0.054	0.045
223	1.143	0.375	0.174
224	1.035	0.676	0.096
225	1.279	1.308	0.216
226	1.309	1.272	0.144
227	1.244	1.594	0.128
228	0.917	0.851	0.119
229	0.942	0.006	0.130
230	0.652	-0.665	0.000
231	0.937	0.437	0.142
232	1.036	0.084	0.214
233	0.718	-0.668	0.000
234	1.085	0.812	0.062
235	0.873	0.016	0.110
236	0.986	-0.157	0.226
237	1.264	0.651	0.155
238	1.114	0.925	0.188
239	0.671	-0.209	0.039
240	1.345	1.581	0.212
241	0.752	-0.547	0.000
242	1.127	1.604	0.109
243	1.392	-0.041	0.334
244	1.891	0.783	0.386
245	1.197	0.939	0.143
246	0.797	0.279	0.115
247	1.170	0.764	0.141
248	1.846	0.159	0.415
249	1.018	0.315	0.172
250	1.325	0.687	0.214
251	1.343	1.406	0.331
252	0.920	1.988	0.181
253	0.845	0.518	0.048
254	1.186	-0.023	0.280
255	0.775	0.626	0.127
256	1.095	1.638	0.278
257	1.181	0.792	0.222
258	1.342	-0.140	0.361
259	1.780	1.418	0.296
260	0.880	0.546	0.242
261	0.441	0.384	0.000
262	0.707	1.340	0.212
263	0.636	-0.227	0.383
264	0.727	1.520	0.063
265	0.627	1.126	0.037
266	0.448	-0.181	0.000
267	0.701	1.182	0.170
268	1.007	1.666	0.130
269	0.537	-1.889	0.119
270	0.541	-2.393	0.119
271	1.322	-0.046	0.025
272	0.661	-1.747	0.142
273	0.693	-2.455	0.110
274	0.941	0.902	0.163
275	0.621	0.406	0.221
276	1.133	0.593	0.074
277	0.861	-0.932	0.050
278	0.536	-0.998	0.000
279	0.695	1.404	0.227
280	0.451	-1.597	0.136
281	0.487	-1.868	0.136
282	0.582	1.023	0.202
283	0.941	1.329	0.262
284	0.627	-1.215	0.120
285	0.479	-0.396	0.120
286	0.496	0.010	0.000
287	1.079	1.034	0.116
288	0.670	-0.509	0.205

Table 17 (cont’d).

Item number	a	b	c
289	0.573	-1.098	0.133
290	0.875	0.955	0.184
291	0.440	2.007	0.006
292	0.717	-1.470	0.115
293	1.309	1.097	0.106
294	0.449	-0.541	0.139
295	0.409	-1.488	0.139
296	0.710	0.259	0.282
297	0.593	-0.813	0.139
298	0.414	-2.427	0.139
299	1.035	1.018	0.181
300	0.591	1.090	0.109
301	0.545	1.571	0.138
302	0.962	1.308	0.030
303	0.737	-1.995	0.101
304	0.706	0.049	0.101
305	1.121	1.172	0.301
306	1.053	1.446	0.291
307	0.835	1.221	0.029
308	0.590	-0.164	0.000
309	1.322	1.567	0.212
310	1.181	1.432	0.058
311	1.127	1.453	0.233
312	1.345	1.150	0.142
313	0.603	0.821	0.263
314	0.711	-0.200	0.308
315	0.547	0.695	0.029
316	0.987	1.359	0.429
317	0.545	1.163	0.142
318	1.058	1.425	0.153
319	0.735	-2.083	0.133
320	0.740	0.640	0.000
321	0.628	-0.594	0.133
322	0.445	-3.375	0.137
323	0.569	-1.032	0.137
324	0.525	0.058	0.137
325	0.767	0.010	0.147
326	0.624	0.237	0.039
327	0.728	1.178	0.084
328	0.804	-0.528	0.143
329	0.616	-0.298	0.000
330	0.634	-0.496	0.000
331	0.757	1.163	0.184
332	1.096	0.844	0.153
333	0.552	0.439	0.148
334	1.454	0.682	0.123
335	0.598	1.470	0.023
336	0.816	1.940	0.257
337	0.575	0.099	0.083
338	0.709	0.446	0.245
339	0.872	2.333	0.017
340	0.734	2.459	0.090
341	0.595	1.869	0.206
342	0.531	1.567	0.302
343	0.652	-0.274	0.000
344	0.588	-0.224	0.000
345	0.672	0.915	0.178
346	0.423	-2.794	0.134
347	0.601	-2.548	0.134
348	0.537	0.682	0.000
349	0.569	0.358	0.279
350	1.306	1.439	0.409
351	0.565	-0.107	0.099
352	0.556	0.359	0.175
353	0.410	-0.168	0.000
354	0.510	-2.375	0.123
355	0.742	0.733	0.182
356	0.651	-1.610	0.000
357	0.529	0.633	0.250
358	0.617	-0.658	0.500
359	1.050	1.099	0.388
360	1.158	1.763	0.197
361	0.772	0.425	0.125
362	0.536	0.486	0.245
363	1.195	1.607	0.221
364	0.895	0.406	0.158
365	1.125	0.817	0.104
366	0.986	0.789	0.254
367	1.180	1.018	0.117

APPENDIX B

Identifying optimal z_c values

The alternative ability estimation procedure utilizes hypothesis-testing, and by necessity requires a critical z-value in order to determine whether the null hypothesis should be retained or rejected. Although the choice of z_c is arbitrary, it stands to reason that some z_c values will lead to more correct decisions—that is, to use the alternative ability estimate when it is more accurate than the conventional ability estimate—than other values of z_c. Further, because ability estimates under ML are characteristically different than those obtained under MAP, it is also possible that the better-functioning z_c values found for the alternative ability estimation procedure concurrent with ML (or, Alt/ML) may be different than those for the alternative ability estimation procedure concurrent with MAP (Alt/MAP). It is then desirable to find the best-functioning, or "optimal" z_c values for the Alt/ML and Alt/MAP procedures. The following provides the empirical basis for which the optimal z_c values were obtained in this study.

This appendix is divided into two sections, with each section further divided into two subsections. The first section discusses how the optimal z_c value was obtained for ability estimation by Alt/ML; the second section discusses how the optimal z_c value was obtained for Alt/MAP. Within each of these sections, two outcome measures are examined: the first is an accuracy measure, and the second is a relative efficiency measure. The two measures are used to provide convergent evidence for the selection of optimal z_c values. A discussion of the two measures follows.

Accuracy measures

Both of the alternative ability estimation procedures are invoked when a decision to reject the null hypothesis is obtained. Thus, the frequency with which the alternative procedures are invoked may be calculated; in addition, the conditional probability that the alternative procedure is more accurate than the conventional procedure, given that the alternative procedure is invoked, may be calculated. To determine whether the alternative procedure is more accurate than the conventional procedure, the absolute difference $\left|\hat{\theta}^* - \theta\right|$ for the alternative ability estimate is compared to the absolute difference $\left|\hat{\theta} - \theta\right|$ for the conventional (ML or MAP) ability estimate $\hat{\theta}$, where θ indicates true ability level. If $\left|\hat{\theta}^* - \theta\right| < \left|\hat{\theta} - \theta\right|$, then the alternative procedure is said to be more accurate than the conventional procedure. If n_{acc} represents the number of times the alternative procedure is more accurate, and n_{inv} represents the number of times the alternative procedure is invoked, then the conditional probability P(accurate|invoked), or P(acc|invoked) is equal to n_{acc}/n_{inv}. If n_{tot} examinees are considered, then the probability that the alternative procedure is invoked for these examinees is P(invoked), and is equal to n_{inv}/n_{tot}.

Simulations were conducted where 500 examinee responses were simulated per ability level $\theta = \{-2, -1, 0, 1, 2\}$ for tests of 25 items in length. A simulation was defined by specific choice of ability estimation procedure (Alt/ML or Alt/MAP) and specific choice of z_c value. The z_c values tested ranged from 0.6 to 1.4 in increments of 0.1; thus, nine z_c values were considered per ability estimation procedure. Maximum FI item selection was used for all simulations. The conditional probability P(acc|invoked) was calculated for each test length i, where $i = \{1, 2, \ldots, 25\}$. Because it was desired to calculate the accuracy of the alternative procedure for all ability levels simultaneously, P(acc|invoked) was not further conditioned on examinee true ability.

Relative efficiency measures

Further evidence for selecting an optimal z_c value was obtained from the second outcome measure, 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). If the alternative ability estimation procedure is more efficient than the conventional estimation procedure, relative efficiency measures should be greater than 1; conversely, if the alternative procedure is less efficient, the measures will be less than 1. Proper selection of z_c should minimize relative efficiency measures that are less than 1 and maximize those measures that are greater than 1. If it is found that this minimization-maximization is impossible, then the use of the alternative ability estimation procedure is not warranted, as it will cause more "harm" than item selection based on conventional ability estimates, and further will not increase efficiency.

As with the accuracy measures, potential z_c values ranged from 0.6 to 1.4 in steps of 0.1. Simulations were conducted for 2500 examinees (500 per true ability level) and for tests of length 5, 10, 15, and 25 items. Maximum FI item selection was used for all simulations. Relative efficiency at true ability θ was computed as the ratio of test information at $i = \{5, 10, 15, 25\}$ items under the alternative procedure, or $I_{ALT}^{(T)}(\theta)$, to the test information at i items under the conventional procedure, or $I_{CONV}^{(T)}(\theta)$. Thus, simulations under ML/Alt and ML were used to identify the optimal z_c for the Alt/ML procedure; likewise, simulations under MAP/Alt and MAP were used to identify the optimal z_c for the Alt/MAP procedure.

Optimal z_c for the Alt/ML procedure

As discussed, two outcomes measures were analyzed in order to determine the optimal z_c value under Alt/ML ability estimation. The first of these is the accuracy measure P(acc|invoked); the second is the relative efficiency measure. The analysis of the accuracy measures is considered first.

Figures 19 and 20 illustrate the accuracy measures for the range of z_c values tested; i.e., 0.6 to 1.4 in increments of 0.1. The accuracy measures are plotted against item administration number. For the group of z_c values shown in Figure 19, the accuracy measures are not constant; rather, they decrease with increasing item administration number. Such a trend is undesirable and therefore these z_c values (0.6, 0.7, and 0.8) are labeled as “unstable.”

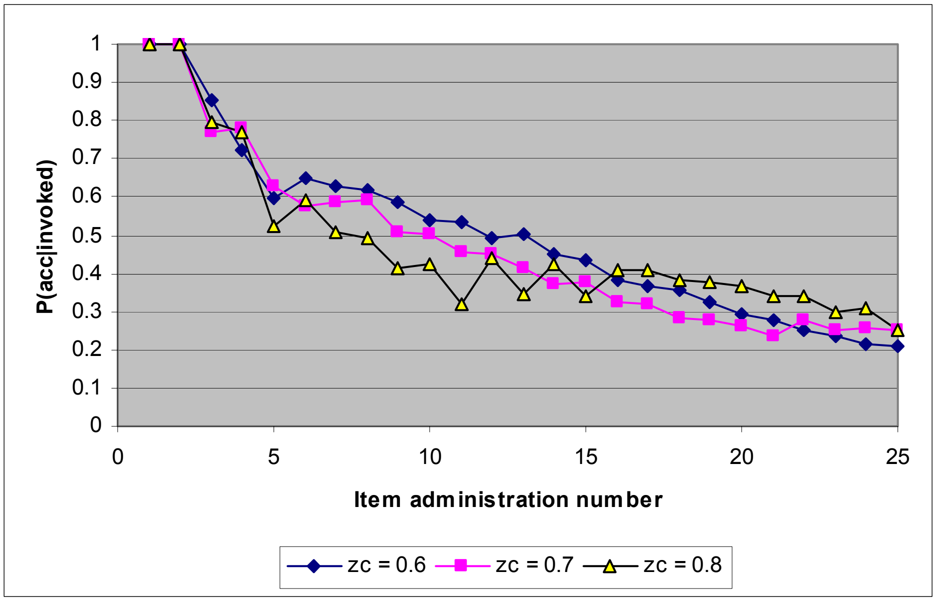

Figure 19. Accuracy measures for Alt/ML ability estimation; unstable set of z_c values.

The sequence of accuracy measures observed for the z_c values in Figure 20 are different from those observed for the unstable z_c values in Figure 19. Most notably, the accuracy measures in Figure 20 stabilize with increasing item administration number. That is, the accuracy measures for this set of z_c values do not decline with increasing test length as they do in Figure 19 (although a slight drop in accuracy is observed for z_c = 0.9 for tests longer than 21 items). Thus, these z_c values are labeled as “stable.”

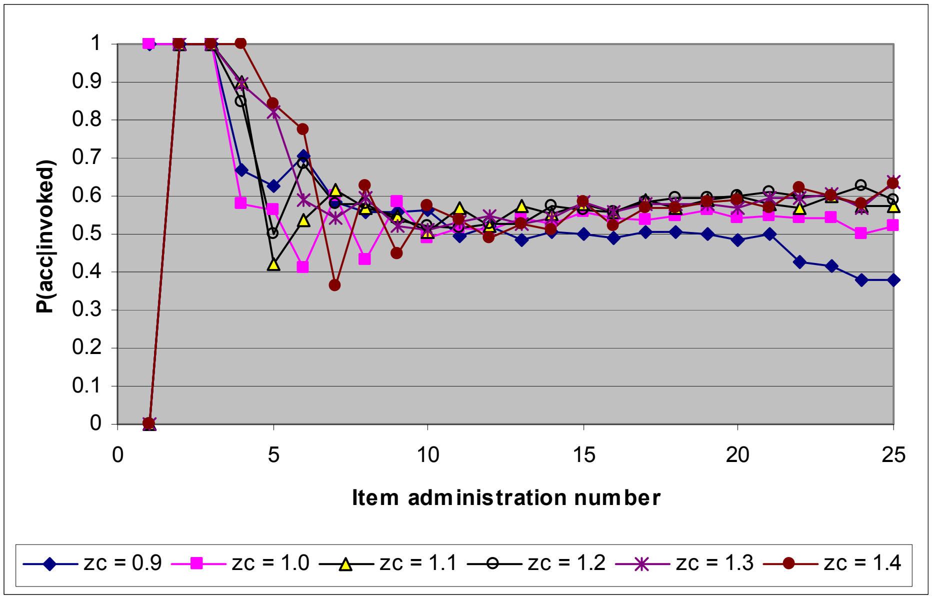

Figure 20. Accuracy measures for Alt/ML ability estimation; stable set of z_c values.

From Figure 20, it appears that higher z_c values are associated with greater accuracy measures in the long run. Unfortunately, higher z_c values necessarily limit the number of times the alternative procedure may be invoked. The probability that the alternative procedure is invoked decreases with increasing z_c, as shown in Figures 21and 22. Whereas the procedure is invoked more frequently for smaller values of z_c, the accuracy measures suffer. Thus, these z_c values 0.6, 0.7, and 0.8 are too liberal. On the other hand, for $z_c \geq 0.9$, the accuracy measures stabilize, although higher z_c values are associated with more conservative tests. Ideally, the optimal z_c value should lead to stable accuracy measures and lead to as many invocations as possible. Based on this criterion, the optimal z_c value for Alt/ML should be no less than (and preferable equal to) 0.9.

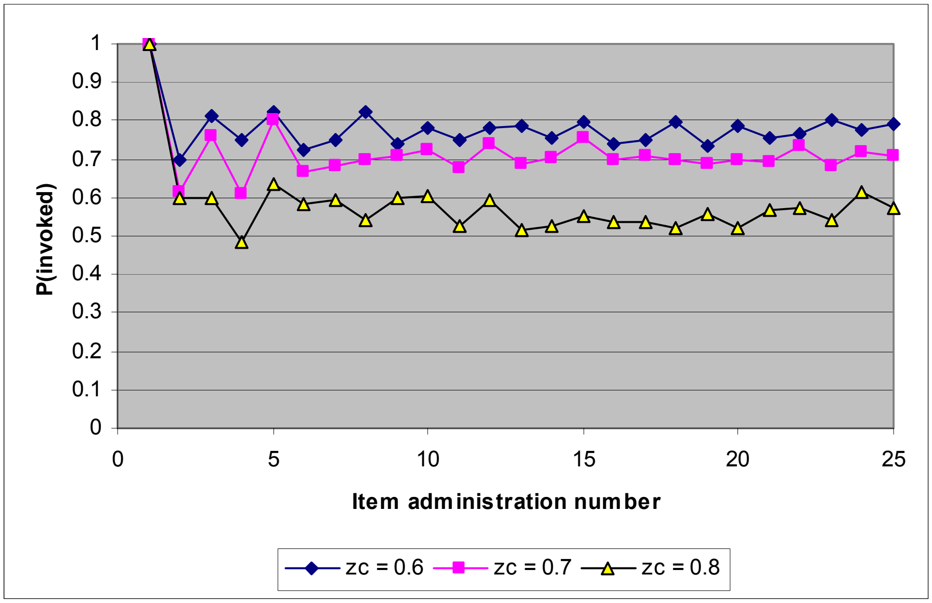

Figure 21. Probability of alternative procedure invocation concurrent with ML estimation; unstable set of z_c values.

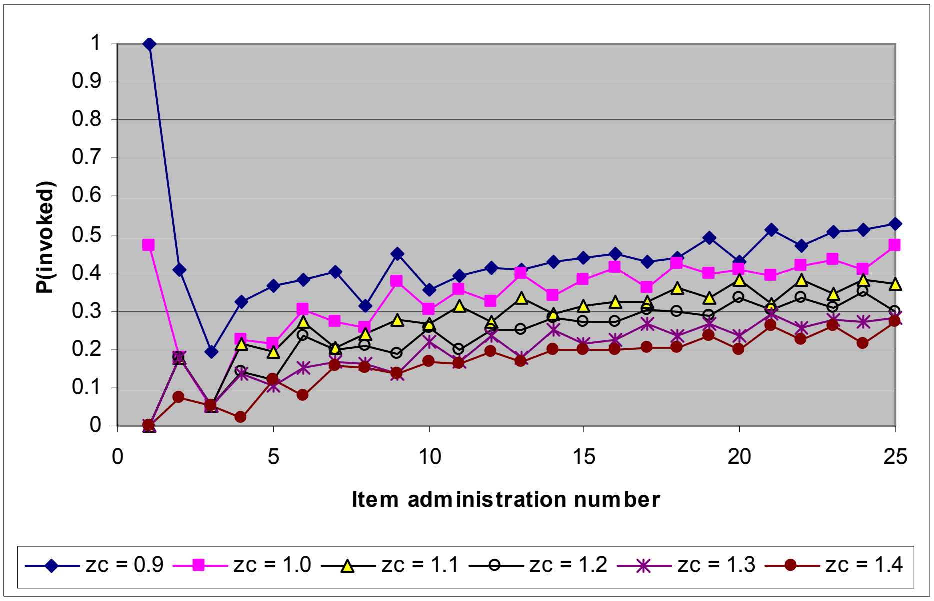

Figure 22. Probability of alternative procedure invocation concurrent with ML estimation; stable set of z_c values.

Additional evidence for the optimal selection of z_c for the Alt/ML procedure comes from an examination of relative efficiency measures. Relative efficiency measures are plotted in Figures 23 and 24, where the five major groups are defined by ability level θ = {-2, -1, 0, 1, 2, and within each group the relative efficiency measures are provided for tests of length i = 5, 10, 15, and 25 items. Figure 23 illustrates the relative efficiency measures when the unstable set of z_c values were chosen. Although the relative efficiency of Alt/ML over the conventional ML is always ≥ 1 for θ = {-1, 0, 1}, it is quite unsatisfactory at the extremes of ability; i.e., θ = {-2, 2}. In these cases, the alternative ability estimation procedures lead to less efficient test administrations, and are clearly at a disadvantage with respect to the conventional ability estimation procedures.

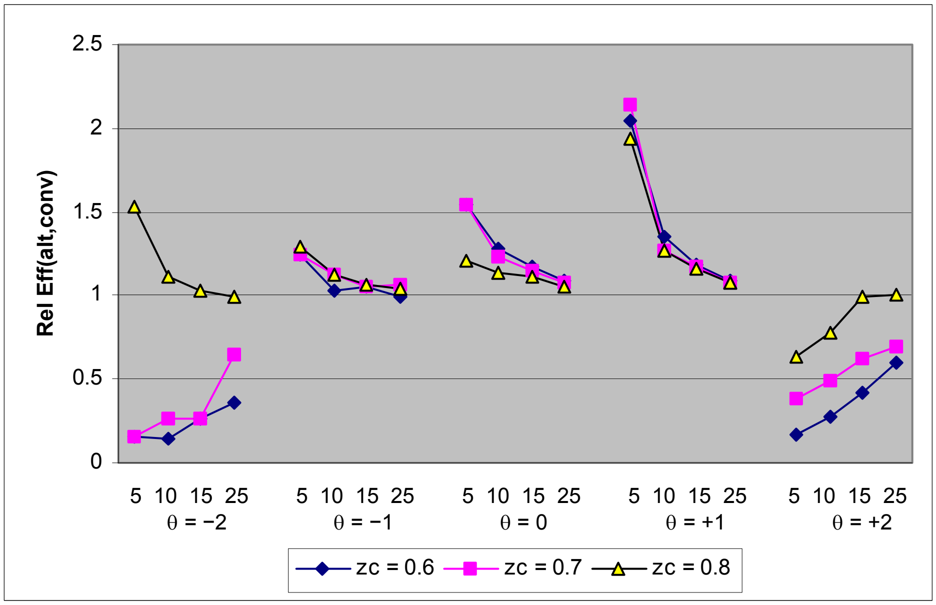

Figure 23. Relative efficiency measures for Alt/ML ability estimation; unstable set of z_c values.

For the set of stable z_c values, the relative efficiency measures are ≥ 1 across all ability levels; i.e., θ = {-2, -1, 0, 1, 2}. However, the conservativeness of the higher z_c values is apparent for $z_c \geq 1.1$, as the relative efficiency measures are nearly equal to 1.0.

It is only for z_c equal to 0.9 or 1.0 that an increase in efficiency is observed for Alt/ML. The increase in efficiency is especially pronounced for z_c = 0.9.

Based on the analysis of the accuracy measures P(acc|invoked), it was concluded that the optimal z_c value for Alt/ML should be no less than 0.9, and preferably equal to it. Analysis of the relative efficiency measures also rules out z_c values less than 0.9; further, it supports the selection of z_c = 0.9 as the optimal z_c value for the Alt/ML ability estimation procedure.

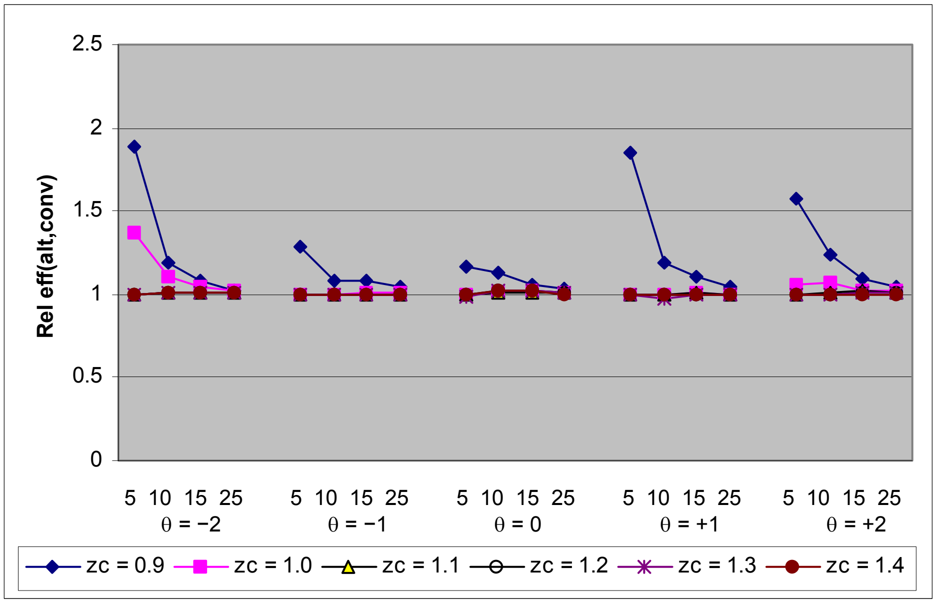

Figure 24. Relative efficiency measures for Alt/ML ability estimation; stable set of z_c values.

Optimal z_c for the Alt/MAP procedure

The analysis of the two outcomes measures, accuracy and relative efficiency, for identifying the optimal z_c value for the Alt/MAP procedure parallels the analysis conducted for the Alt/ML procedure. Again, the analysis of the accuracy measures is considered first.

Figures 25 and 26 illustrate the accuracy measures observed for the Alt/MAP simulations where test z_c values 0.6 through 1.4 in increments of 0.1 were chosen. As shown in Figure 25, z_c values ≤ 1.0 led to unstable accuracy measures, although the pattern of instability differs from that observed in the Alt/ML case (as shown in Figure 19). For z_c values equal to 0.8, 0.9, and 1.0, Figure 25 shows a somewhat erratic pattern in the accuracy measures as test length increases. For the smallest z_c values, the pattern appears stable but it is quite low in magnitude, with P(acc|invoked) equal to about 0.2. The z_c values shown in Figure 25 are thus not likely to be optimal for Alt/MAP.

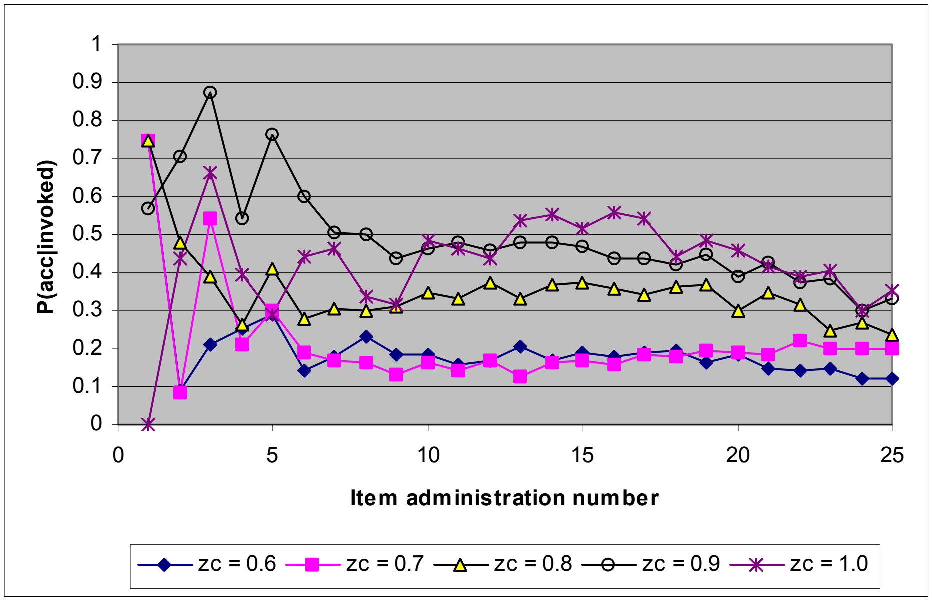

Figure 25. Accuracy measures for Alt/MAP ability estimation; unstable set of z_c values.

In contrast, the z_c values shown in Figure 26 are stable for increasing test lengths, and the accuracy measures are higher. An anomaly appears for $z_c = 1.4$ and a test length equal to 4 items, where the accuracy measure suddenly drops to zero, only to return to approximately 0.5 for the 5th item. Of the four z_c values shown in Figure 26, $z_c = 1.3$ leads to the most desirable sequence of accuracy measures, in that the sequence possesses no anomalies and never decreases with increasing item administration number. Based on this analysis, the optimal choice of z_c for the Alt/MAP procedure should not be less than 1.1 (and preferably equal to it).

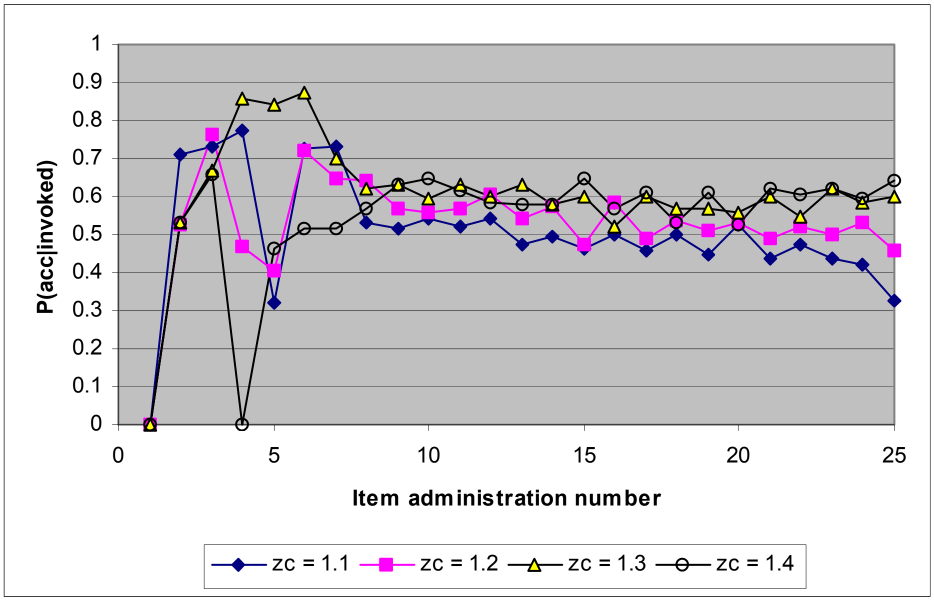

Figure 26. Accuracy measures for Alt/MAP ability estimation; stable set of z_c values.

As noted in the identification of the optimal z_c for Alt/ML, increasing z_c values lead to decreasing frequencies for invocation of the alternative ability estimation procedure. The probability that the procedure is invoked for the unstable set of z_c values is illustrated in Figure 27; the corresponding plots for the stable set of z_c values are shown in Figure 28.

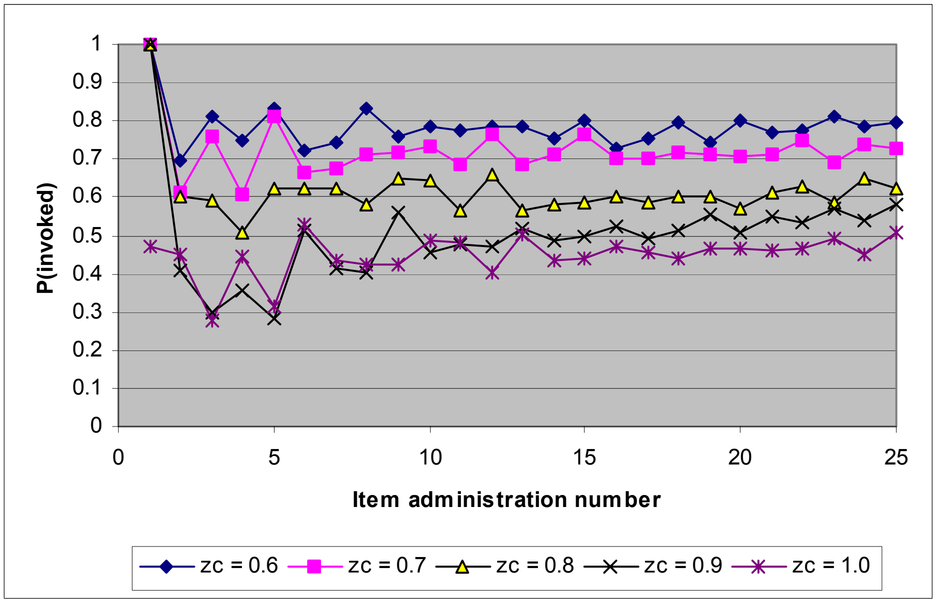

Figure 27. Probability of alternative procedure invocation concurrent with MAP estimation; unstable set of z_c values.

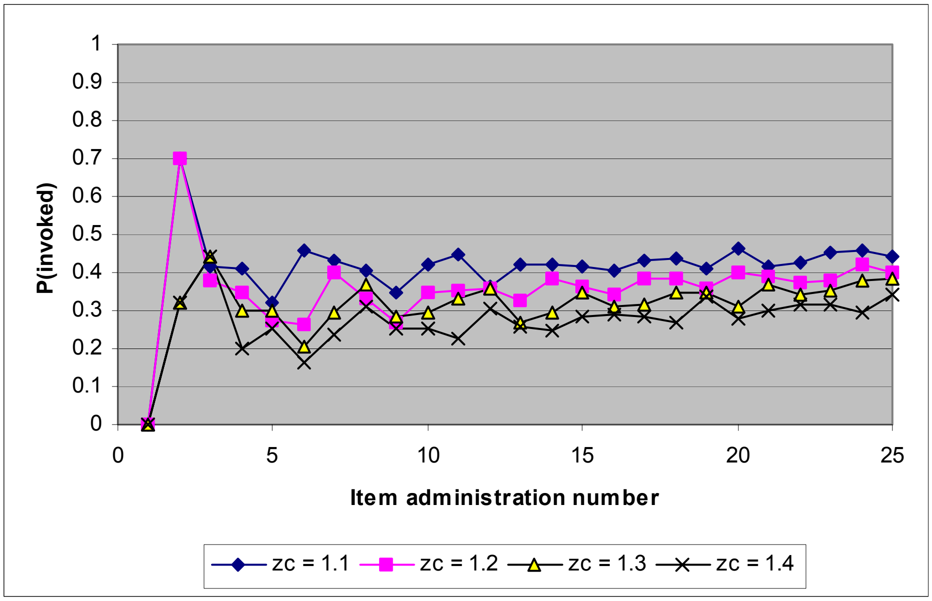

Figure 28. Probability of alternative procedure invocation concurrent with MAP estimation; stable set of z_c values.

The relative efficiency measures for Alt/MAP over conventional MAP estimation are provided in Figures 29 through 32. Figure 29 shows the relative efficiency measures for the unstable set of z_c values; Figure 30 is a rescaling of Figure 29 to show relative efficiency measures close to 1. For the very smallest values of z_c (0.6 and 0.7), the relative efficiency is always less than 1 for ability levels at -2 and 2. For the remainder of z_c values in this set, the relative efficiency measures drop substantially below 1. Thus, none of these unstable z_c values should be considered as optimal.

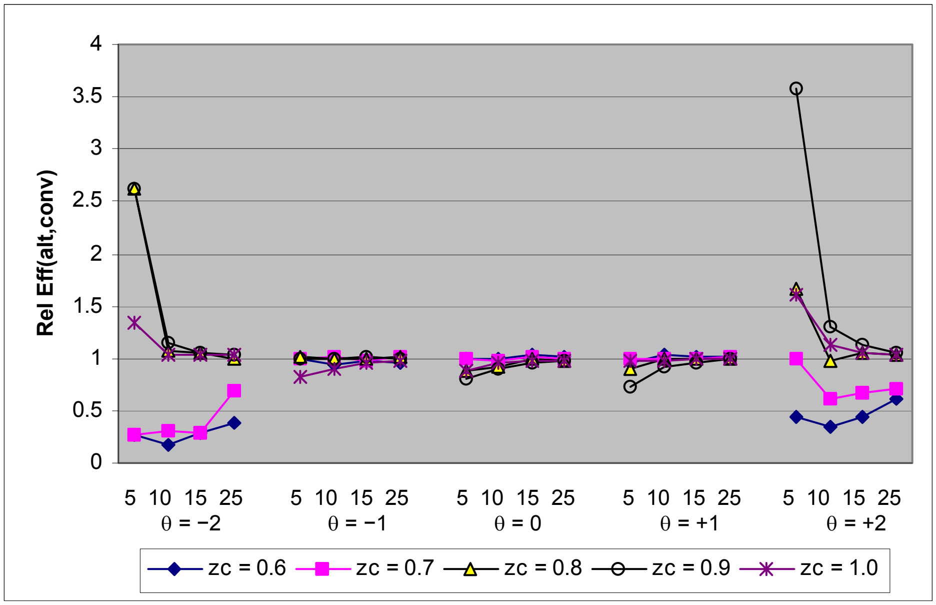

Figure 29. Relative efficiency measures for Alt/MAP ability estimation; unstable set of z_c values.

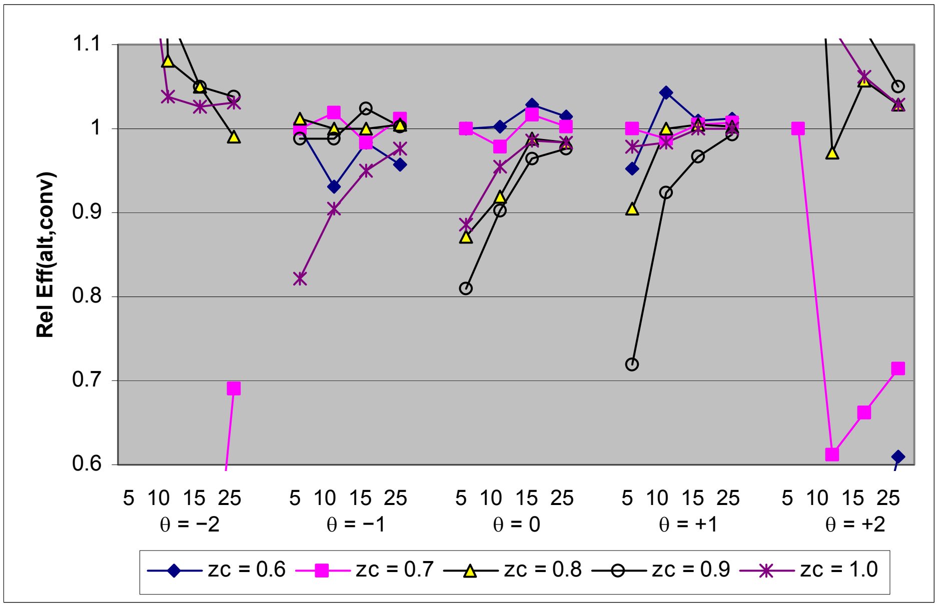

Figure 30. Relative efficiency measures (rescaled) for Alt/MAP ability estimation; unstable set of z_c values.

The stable set of z_c values, as illustrated in Figures 31 and 32, clearly show improved relative efficiency measures as compared to the unstable set of z_c values. Examining Figure 31 suggests that either $z_c = 1.2$ or $z_c = 1.3$ could be selected as an optimal z_c value; examining Figure 32 shows that inefficiencies introduced by Alt/MAP (with respect to conventional MAP) are minimized for $z_c = 1.3$; these inefficiencies occur when $\theta = \{-1, 0, 1\}$. Further, although $z_c = 1.2$ leads to greater relative efficiency for $\theta = 2$, it is less efficient than $z_c = 1.3$ for $\theta = -2$. Thus, $z_c = 1.3$ is recommended based on the relative efficiency measures.

In sum, for Alt/MAP, the accuracy measures suggested that z_c should be no less than 1.1; in addition, the best sequence of the accuracy measure versus item administration number was obtained for $z_c = 1.3$ An analysis of the relative efficiency measures also suggested that $z_c = 1.3$ be selected as an optimal choice of z_c.

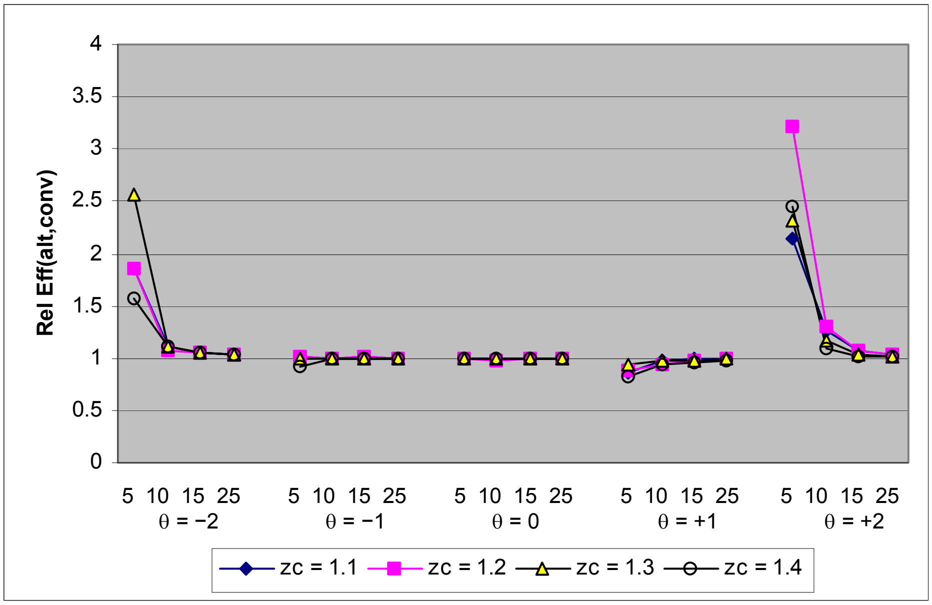

Figure 31. Relative efficiency measures for Alt/MAP ability estimation; stable set of z_c values.

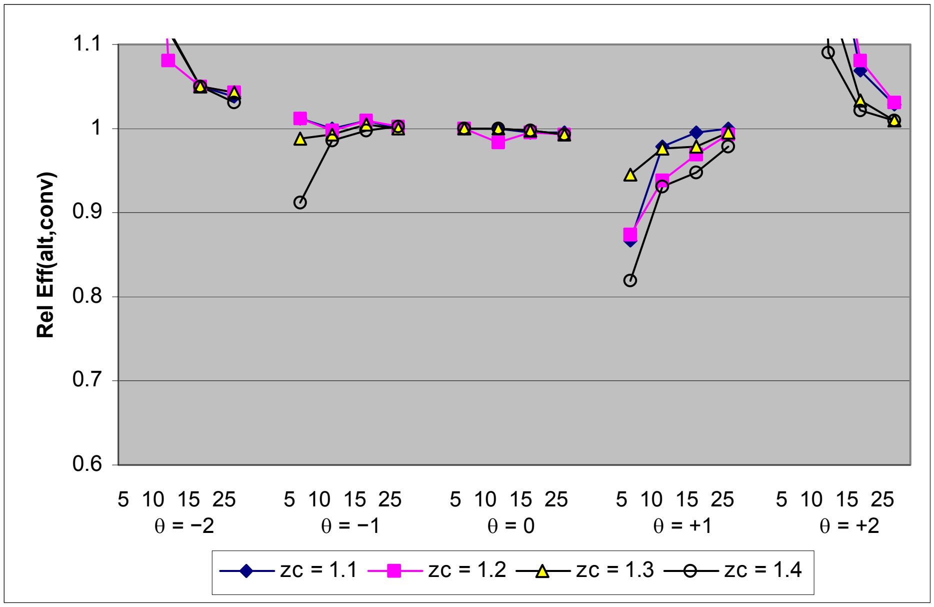

Figure 32. Relative efficiency measures (rescaled) for Alt/MAP ability estimation; stable set of z_c values.

APPENDIX C

SimCAT: A program for a CAT simulation environment

The SimCAT program is divided into two sections: the first is the main code for executing the program, the second contains the SAS macro code.

SimCAT Main

```
options notes;
options mprint;

********************************************************************;
* SimCAT
*
* SAS program for generating IRT mixed-model data;
*                       estimating ability;
*                         calculating item information;
*                         administering a CAT;
*                         estimating & utilizing CAT exposure control
*
* Written by Alexander Weissman
* April 12, 2002;
********************************************************************;

/*  This program has three modes of operation:

      [Note that < &macro_variable > denotes file specified by macro variable]

      Mode 1:  Simulate IRT responses under mixed-models
                   Estimate ability from these responses (unidimensional only)
             Calculate item information

                   INPUT FILES/DATASETS:
                              <&parmfile>       Item parameters
                              <&seedfile>       Random number seed file

                   OUTPUT FILES/DATASETS:
                              <&outfile>        Simulated item responses
                              GEN_DATA          SAS dataset of item responses
                              EAP                     SAS dataset of EAP estimates
                              INFOFN            SAS dataset of item information functions
```

```
  Mode 2:  Administer a CAT by simulating examinee abilities
               from a known distribution

             INPUT FILES/DATASETS:
                         <&parmfile>       Item parameters
                         <&seedfile>       Random number seed file

               OUTPUT FILES/DATASETS:
                         CAT                     SAS dataset of CAT administration

  Mode 3:  Administer a CAT interactively, specifying examinee
                  ability in advance, with the option of producing
                    provisional ability sequence plots

                INPUT FILES/DATASETS:
                         <&parmfile>       Item parameters
                         <&seedfile>       Random number seed file

                OUTPUT FILES/DATASETS:
                         CAT                     SAS dataset of CAT administration

  Mode 4:  Estimate Sympson-Hetter exposure control parameters, conditioning
         on ability if desired


  The macro variable &MODE = {1, 2, 3, 4} selects the mode of operation */

%let mode = 3;

  /*******************************************************************/
  /*  Control information */
  /*******************************************************************/

  /*  Input & output files *************************/

  /*  Assign working directory */
  filename wrkdir 'c:\dimension\crdf';
  /*  Identify input file for item parameters */
  %let parmfile='itemparms.par';

  /*  Choose output filename for IRT simulated responses */
  %let outfile='responses.dat';

  /*  Identify random number seed file */
  /*  Random number seeds should be updated after each iteration
      of program for simulation studies */
  %let seedfile='gen.rsd';

  /*  Identify input file for exposure control parameters */
  %let expconin = 'itemparms.exp';

  /*  Identify output file for exposure control parameters */
  %let expconout = 'itemparms.exp';

  /*  Identify macro file */
  %let macrofile='simcat_macros.sas';

  /*  Characteristics of data simulation ******************************/

  /*  Define the size and covariance matrix for the ability space */
  /*  Note that this matrix should conform to the dimensions specified by &ndims */
  /*  Unidimensional IRT models should have ndims = 1 and sigmatrix = {1} */
%let ndims = 1;
%let sigmatrix = {1};

  /*  Identify parameters for Fleishman's (1978) power transformation
        on ability distributions.  Useful for generating non-normal distributions
      of ability */

%let transb=1;
  %let transc=0;
  %let transd=0;

  /*  Declare dimensions of item parameter data spaces */
```

```
      /*  Variable &nparms should be set equal to the maximum
          number of parameters across all items */
      %let nparms=3;

      /*  Maximum number of response categories for items */
      %let maxcat=2;

      /*  Set scaling parameter for IRT probability functions */
    %let d=1.702;


/*  Item types by IRT model saved in *.par file, listed here for reference */
/*
    Itemid:
         item identification field

      Model types:

    di = dichotomous 3PL (unidimensional)
    md = dichotomous 3PL (multidimensional)
      gr = graded response
    pc = partial credit
    gpc = generalized partial credit

    Item parameters are saved in file designated by &parmfile
      and formatted as follows:

     -- All items --
       Col1                      Col2
      <model type>    <number of categories ={2, 3, ... }>

    -- Dichotomous items --

       Col3  Col4  Col5
        <a>   <b>   <c>

    -- Multidimensional dichotomous items --
```

```
     Col3  Col4  Col5 ... Col_k  Col_k+1  Col_k+2
       <a1>  <a2>  <a3> ... <ap>   <beta_0> <c>


     where a1...ap are the discrimination parameters for the p dimensions
     beta_0 is the intercept parameter (related to item difficulty)
     c is the pseudo-guessing parameter

    -- Graded response --

     Col3  Col4  Col5 ... Col_k
     <a>   <b1>  <b2> ... <bj>

       where a is the discrimination parameter, and b1...bj are the
       item category thresholds

*/

    /* Control settings for ability estimator*****************/

    /*  Set number of quadrature points */
    %let nqpt = 60;

    /*  Set number of fine-scaled quadrature points */
    %let qfine = 20;

    /*  Set upper bound for ability estimation, theta scale */
    %let ub = 4;

    /*  Quadrature increment */
    %let xinc=( (2*&ub)/(&nqpt-1) );

    /*  Convergence criterion for Newton-Raphson */
    %let crit = 0.001;

    /*  Maximum number of iterations for Newton-Raphson routine */
    %let newton_max = 20;
```

```
      /*  Set mean and variance for prior distribution */
      %let mean = 0;
      %let sigma2 = 1;

      /*  Miscellaneous settings ****************************/

      %let missing = 9;

      /*  Default settings for item exposure counters
            This information is used by the program even if exposure control
          is set to OFF */

      %let bins = 1;

      /*  Load macros ***************************************/
      %include wrkdir(&macrofile);

%macro model;

/*********************;
/* Simulate IRT data  ***********************/
/*                                                                */

      /*  Declare number of items and subjects */
      %let nitems=36;
      %let nsubj=100;

      /*  Simulate data */
      %seeds;
      %thetagen;
      %itempar;
      %mergegen;
      %gendata;
      run;

      /*  Estimate ability */
      %quad;
      %mergeabl;
      %likehood;
      %eapest;
      run;

      /*  Compute item information functions */
      %quad;
      %infofn;
      run;

      /*  Compute theta_max and I(theta_max) */
      %itempar;
      %infomax;
      run;

%mend mode1;

%macro mode2;

/*********************;
/* CAT Administration ***********************/
/*                                                                */

      /*  Number of examinees */
      %let nsubj=10;

      /*  Estimation method 1=ML, 2=MAP, 3=EAP */
      %let est=2;

      /*  Item selection method 1 = FI, 2 = FII */
      %let itemsel=2;

            /*  Z-value for FII interval width */
            %let zfii = 2;

      /*  Exposure control options 0 = OFF, 1 = ON */
      %let expcon = 0;

      /*  Type of CAT:  1 = fixed-length, 2 = variable-length */
      %let cattype=1;

            /*  All of the following may be declared, but only some will be active
              depending on value of &cattype */

            /*  For fixed-length CAT
                  Length of CAT administration, in number of items */
                  %let catlength=20;

            /*  For variable-length CAT
                  Precision of measurement (standard error) */
                  %let catprec=0.4;

            /*  Limit for variable-length CAT, in number of items */
                  %let catlimit=30;

      /*  Call CAT executive macro */
      %cat;

%mend mode2;

%macro mode3;

*****************************;
*  Interactive CAT                    ***********************;
*                                                                         *;
*                                                                         *;

      /*  Specify input file for theta values */
      %let ifile='ithetas3.dat';

      /*  Choose number of replications for each theta value */
      %let nreps = 1000;

      /*  Estimation method 1=ML, 2=MAP, 3=EAP,
                            4=ML grid search, 5=MAP grid search
                            6=GSS                              */
      %let est=1;
```

```
/*  Ability estimation experimental modes:
    0 = conventional, 1 = alternative, 2 = Hybrid, 3 = True */
%let expselect = 0;

      /*  Critical z-value for alternative mode */
      %let ztol = 1;

/*  Information measure output: 1 = At true, 2 = At provisional */
%let infomeas = 1;

/*  Item selection method 1 = FI (exact), 2 = FII (exact) */
%let itemsel=1;

      /*  If FII, specify interval width */
      %let zfii = 2;

/*  Exposure control options 0 = OFF, 1 = ON */
%let expcon = 0;

/*  Type of CAT:  1 = fixed-length, 2 = variable-length */
%let cattype=1;

      /*  All of the following may be declared, but only some will be active
        depending on value of &cattype */

      /*  For fixed-length CAT
            Length of CAT administration, in number of items */
            %let catlength=25;

      /*  For variable-length CAT
            Precision of measurement (standard error) */
            %let catprec=0.4;

      /*  Limit for variable-length CAT, in number of items */
            %let catlimit=30;

/*  Graphing options 0 = no graphs, 1 = graphs */
```

```
      %let graphs=0;

      %icat;
      %graphs;
    quit;

%mend mode3;

%macro mode4;

/*************************************;
/* Estimate exposure control parameters ***********/
/* for a CAT                                                        */

      /* Method:  conditional or unconditional on ability
         1 = conditional, 0 = unconditional */
      %let cond = 1;

            /*  If conditional Sympson-Hetter, then specify number of
                  ability groups.  This setting has no effect if
                  unconditional S-H is used.  */
                  %let bins = 2;

      /*  Target exposure probability for Sympson-Hetter technique */
      %let target=0.40;

      /*  Tolerance for maximum exposure rate (e.g., 0.05 = 5%) */
      %let tolerance = 0.10;

      /*  Maximum number of iterations */
      %let maxiter = 1;

      /*  Number of examinees */
      %let nsubj=100;

      /*  Estimation method 1=ML, 2=MAP, 3=EAP */
      %let est=1;
      /*  Item selection method 1 = FI, 2 = FIP */
      %let itemsel=1;

      /*  Exposure control always set to 1 (ON) */
      %let expcon = 1;

      /*  Type of CAT:  1 = fixed-length, 2 = variable-length */
      %let cattype=1;

            /*  All of the following may be declared, but only some will be active
              depending on value of &cattype */

            /*  For fixed-length CAT
                  Length of CAT administration, in number of items */
                  %let catlength=20;

            /*  For variable-length CAT
                  Precision of measurement (standard error) */
                  %let catprec=0.4;

            /*  Limit for variable-length CAT, in number of items */
                  %let catlimit=30;

      /*  Call EXPCAT (Exposure control CAT) executive macro */
      %expcat;

%mend mode4;

%macro execute;

      /*  Declare global variables */
      %global npool nitems est;

      /*  Direct program functionality */

      %if &mode=1 %then %do;
            %mode1;
            %end;
      %if &mode=2 %then %do;
            %mode2;
            %end;

      %if &mode=3 %then %do;
            %mode3;
            %end;

      %if &mode=4 %then %do;
            %mode4;
            %end;

%mend execute;

%execute;
```

SimCAT Macros

```
/**********************************************/
/*                                                                  */
/*  SimCAT MACROS                                         */
/*                                                                  */
/*  Written by Alexander Weissman                   */
/*  April 12, 2002                                              */
/**********************************************/

/*  Start date:  September 15, 2001 */
/*  Modified:  April 12, 2002 */

/*  MACRO LIST */

/*  CAT:       Highest-level module for executing a CAT administration */
/*  EAPEST:    Estimate ability for examinees using EAP */
/*  GENDATA:   Generates item responses */
/*  INFOFN:    Computes information function for items at quadrature points */
/*  INFOMAX:   Computes theta_max and info_max for items */
/*  ITEMPAR:   Loads item model type, item parameters, and number of categories for item */
/*  LIKEHOOD:  Computes likelihood function given examinee response pattern */
/*  MERGEGEN:  Merges the item model type, parameters, and number of categories with
               ability matrix generated in THETAGEN */
/*  MERGEABL:  Merge datasets for ability estimation */
/*  QUAD:      Configure quadrature points and weights */
/*  SEEDS:     Loads random number seeds from file */
/*  THETAGEN:  Generates multivariate latent ability space, where distributions
               may normal or non-normal */


/*  DI:  Calculates response probabilities under 1-D 3PL model */
/*  GPC:  Calculates response probabilities under 1-D generalized partial-credit model */
/*  GR:  Calculates response probabilities under 1-D graded response model */
/*  MD:  Calculates response probabilities under multidimensional 3PL model */
/*  PC:  Calculates response probabilities under 1-D partial-credit model */
```

```
%macro seeds;
/******************************************************************************;
/*  SUBROUTINE SEED  */
/*  INPUT DATASETS:  <FILE:  &seedfile>
    OUTPUT DATASETS:  SEEDS;
    OUTPUT MACRO VARIABLES:  &SEEDMVN, &SEEDRESP

    &SEEDMVN is the random number seed for creating the
             multivariate normal deviates in subroutine
             MVNTHETAS

      &SEEDRESP is the random number seed for creating the
             item responses in subroutine GENERATE
*/

/*  Load the random number seed from file, place it in
macro variable &newseed using CALL SYMPUT */

data seeds;
      infile wrkdir(&seedfile);
      input value1 value2;
      call symput('seedmvn',value1);
      call symput('seedresp',value2);
run;
%mend seeds;

%macro thetagen;
*******************************************************************;

/*SUBROUTINE MVNTHETAS

  INPUT DATASETS:  NONE
  OUTPUT DATASETS:  THETAS


  Simulates multivariate normal observations:
```

```
  The following IML routine creates a SAS data set THETAS containing n observations
  simulated from a multivariate normal distribution with mean mu and covariance matrix Sigma. */
proc iml;
  n=&nsubj;
  seed=&seedmvn;
  mu=repeat(0,&ndims,1);  /* create column vector of zeros, with # rows = &ndims */
  Sigma=&sigmatrix;
  p=nrow(Sigma);
  z=normal(repeat(seed,n,p));       /*creates a nxp matrix z of random normal deviates */
  a=(root(Sigma));                  /* compute square root of Sigma matrix such that aa` = Sigma */
  b=repeat(mu`,n,1);                /* create a nxp matrix b of means */
  y=(z*a)+b;                        /* create a nxp matrix y of multivariate normal data */
  create thetas from y;       /* create a SAS dataset with the data */
  append from y;
  close thetas;
quit;

/*  Rename the generic IML "col" variable names to "theta" */
data thetas;
      set thetas;
      rename col1-col&ndims=theta1-theta&ndims;
run;

******************************************************************;

/*  SUBROUTINE TRANSFORM

    INPUT DATASETS:  THETAS
    OUTPUT DATASETS:  THETAS */

/*  This subroutine transforms the MVN thetas to non-normal
(marginal) distributions, based on Fleishman's (1978) power trans-
formation.

Y = a + bX + cX**2 + dX**3

For mean-centered data, constrain a & c such that a = -c.
```

```
Values of a, b, and d may be chosen for desired skewness and
kurtosis.

*/
*****************************************************************;

data thetas;
      set thetas;
      array theta(*) theta1-theta&ndims;

      /*  Set values of b, c, and d for each dimension */
    /*  For no transformation, let b = 1, c = 0, d = 0 */

    b=&transb;
    c=&transc;
      d=&transd;
      a=-c;

      do i=1 to &ndims;
      theta(i) = a + b*theta(i) + c*(theta(i)**2) + d*(theta(i)**3);
      end;

      drop a b c d;
run;
%mend thetagen;

%macro itempar;

*******************************************************************;
/*  SUBROUTINE ITEMPAR

    INPUT DATASETS:  <FILE identified by &parmfile>
    OUTPUT DATASETS  ITEM_PAR_FULL
                               ITEM_PAR
                               MODEL_TYPE
                               NCAT_INFO
;
*READ ITEM Parameters;
```

```
*/
*******************************************************************;

/*  Step 1 of 3:  Extract item parameter information */

/*%global npool nitems;*/

data item_par_full;
  infile wrkdir(&parmfile) missover;
  input itemid $ model $ ncat x1-x&nparms;
  call symput('npool',_n_);  /*  Determine size of item pool */
run;

%let npool=%eval(&npool);

data item_par;
   set item_par_full;
*if reading in parameters for the 1P or 2P models need to set c=0;
*  x3=0;
   array y{*} x1-x&nparms;
   keep p;
   do j=1 to &nparms;
     p=y{j};
     output;
   end;
run;

/*  Create a row vector with item parameters as elements */
proc transpose data=item_par out=item_par prefix=p;
      var p;
run;

/*  Step 2 of 3:  Extract model type information */

data model_type;
      set item_par_full;
      keep model;
run;
```

```
proc transpose data=model_type out=model_type prefix=model;
      var model;
run;

/*  Step 3 of 3:  Extract number of categories information */

data ncat_info;
     set item_par_full;
       keep ncat;
run;

proc transpose data=ncat_info out=ncat_info prefix=ncat;
     var ncat;
run;

%mend itempar;

%macro gr;
*******************************************************************;
/*  SUBROUTINE GR

    Computes probability functions for graded response
    model
*/
*******************************************************************;

    do;
    if resp =  (ncat[j] - 1) then
       xx=1/(1+exp(-&d*p{j,1}*(theta1-p{j,resp+1})));
    else if resp=0 then
       xx=1-1/(1+exp(-&d*p{j,1}*(theta1-p{j,2})));
    else
       xx=1/(1+exp(-&d*p{j,1}*(theta1-p{j,resp+1})))
           -1/(1+exp(-&d*p{j,1}*(theta1-p{j,resp+2})));
      end;
%mend gr;
```

```
%macro di;
*******************************************************************;
/*  SUBROUTINE DI

    Computes probability functions for unidimensional
    3PL dichotomously scored model
*/
*******************************************************************;
     do;
             xx=.;
       xx=p{j,3}+(1-p{j,3})/(1+exp(-&d*p{j,1}*(theta1-p{j,2})));
             if resp=0 then xx=1-xx;
             end;

%mend di;

%macro md;
*******************************************************************;
/*  SUBROUTINE MD

    Computes probability functions for multidimensional
dichotomously scored model with pseudo-guessing parameter
*/

/*  Parameters in parameter list are in the following sequence:
    a1, a2, ... ,ap, beta_0, c

    where a1...ap are the discrimination parameters for the p dimensions
    beta_0 is the intercept parameter (related to item difficulty)
    c is the pseudo-guessing parameter */
*******************************************************************;

      do;

      /*  Step 1:  Assign the beta_0 value to the argument */
      argmnt = p{j,%eval(&ndims+1)};

      /*  Step 2:  Accumulate the terms with discrimination parameters */
      do tht = 1 to &ndims;
           argmnt = argmnt + p{j,tht}*theta(tht);
        end;

        /*  Step 3:  Identify c parameter */
        cparm = p{j,%eval(&ndims+2)};

      xx = cparm + (1 - cparm)/(1 + exp(-&d*( argmnt )));
      if resp=0 then xx=1-xx;
      end;
%mend md;


%macro mergegen;
*******************************************************************;
/*  SUBROUTINE MISC

    Merge item model and number of categories information

    INPUT DATASETS:  MODEL_TYPE
                                NCAT_INFO
    OUTPUT DATASET:  MISC
*/
*******************************************************************;

data misc;
set model_type;
set ncat_info;
run;


******************************************************************;
/*  SUBROUTINE MERGEPARMS

INPUT DATASETS:  ITEM_PAR, THETAS, MISC
OUTPUT DATASETS:  PARMS

This subroutine creates a matrix where the row vector of item
```

```
parameters is repeated &nsubj times alongside a submatrix
consisting of the generated theta values theta_1 .... theta_p

*/
******************************************************************;

data parms;
one=1;    /*  set pointer to one */
merge thetas;
set item_par point=one;
set misc point=one;
run;
%mend mergegen;

%macro gendata;
*******************************************************************;
/*  SUBROUTINE GENDATA

    Simulates IRT data according to specified models

    INPUT DATASETS:  PARMS
    INPUT MACRO VARIABLES:  &NDIMS, &NITEMS, &nparms,
                                          &nparms, &SEEDRESP
    OUTPUT DATASETS:  GEN_DATA
                      SEEDS
*/
******************************************************************;
data gen_data seeds (keep=seed);

      set parms;
      array theta{&ndims} theta1-theta&ndims;
      array p{&nitems,&nparms} p1-p%eval(&nitems * &nparms);
      array ix{&nitems} ix1-ix&nitems;
      array ncat{&nitems} ncat1-ncat&nitems;
      array model{&nitems} model1-model&nitems;
      array cumprob{*} cumprob1-cumprob%eval(&maxcat);

      retain seed &seedresp;
```

```
/*  All unidimensional models use theta1 as the latent variable */
/*  Multidimensional models may use theta1 through theta&ndims */

do j=1 to &nitems;

      do k=1 to &maxcat;
         cumprob[k]=.;
         end;

      do resp=0 to ncat{j}-1;
                  if model[j]="di" then do;
            %di;
                        end;
                  else if model[j]="gr" then do;
                        %gr;
                        end;
                  else if model[j]="md" then do;
                        %md;
                        end;
            if resp=0 then cumprob{1}=xx;
          else cumprob{resp+1}=xx+cumprob{resp};
          end;

            call ranuni(seed,r01);

          do resp=1 to ncat{j}-1;
            if resp=1 and (1-r01) <= cumprob{resp} then
            ix{j}=0;
          else if (1-r01) > cumprob{resp} and (1-r01) <= cumprob{resp+1} then
              ix{j}=resp;
        end;

end;

            /*file wrkdir(&outfile);
      put (ix1-ix20) (1.);*/
```

```
                  run;

%mend gendata;

%macro quad;
*******************************************************************;
/*  SUBROUTINE QUAD

    Configure quadrature points and prior weights

    INPUT DATASETS:  <NONE>
    INPUT MACRO VARIABLES:  &NQPT, &SIGMA, &MEAN
    OUTPUT DATASETS:  QUAD
*/
*******************************************************************;

data quad;
array qpt{&nqpt};
array prior{&nqpt};

/*  Compute increment for quadrature points */
xinc=&xinc;

/*  Assign values for quadrature points, store in qpt{} array */

do j=1 to &nqpt;
        qpt{j}=-&ub+(j-1)*xinc;
end;

arg=2*&sigma2;
pi = 3.1415;
total=0;

/*  Determine ordinates of normal distribution at quadrature points */
/*  If ability estimation is by ML, assign a uniform prior, else assign
    a normal prior */

do j=1 to &nqpt;
```

```
              if &est=1 or &est=4 then do;
                        prior{j}=1;
                    end;
              else do;
                  prior{j}=(1/sqrt(arg*pi))*exp(-(qpt{j}-&mean)**2/arg);
                        end;

end;

run;


%mend quad;

%macro quadgrid;
********************************************************************;
/*  SUBROUTINE QUADGRID

    Configure quadrature points and prior weights for fine-scaled
      quadrature grid

    INPUT DATASETS:  <NONE>
    INPUT MACRO VARIABLES:  &NQPT, &SIGMA, &MEAN, &QFINE
    OUTPUT DATASETS:  QUADGRID
*/
********************************************************************;

%let gridnums = %eval(&nqpt * &qfine);

data quadgrid;
      array qgrid{&nqpt,&qfine}  qgrid1-qgrid&gridnums; /* Quad points */
      array pgrid{&nqpt,&qfine}  pgrid1-pgrid&gridnums; /* Priors */

      /*  Coarse-scaled increment */
      xinc = &xinc;

      /*  Compute fine-scaled increment for quadrature points */
    gridinc=%sysevalf(&xinc/&qfine);
```

```
      /*  Assign values for quadrature points, store in qpt{} array */

      do j=1 to &nqpt;
            do k=1 to &qfine;
            qgrid{j,k}=-&ub + (j-1)*xinc + (k-1)*gridinc;
            end;
      end;

      arg=2*&sigma2;
      pi = 3.14159;
      total=0;

      /*  Determine ordinates of normal distribution at quadrature points */

      do j=1 to &nqpt;
            do k=1 to &qfine;
                  if &est=1 or &est=4 then do;
                        pgrid{j,k}=1;
                        end;
                  else do;
                  pgrid{j,k}=(1/sqrt(arg*pi))*exp(-(qgrid{j,k}-&mean)**2/arg);
                        end;

            end;
      end;


      run;

%mend quadgrid;


%macro mergeabl;
*******************************************************************;
/*  SUBROUTINE MERGEABL

    Merge datasets for ability estimation
    INPUT DATASETS:  GEN_DATA, ITEM_PAR, MISC, QUAD
    INPUT MACRO VARIABLES:  &NDIMS
    OUTPUT DATASETS:  MERGEABL
*/
********************************************************************;

/*  Merge datasets for ability estimation */

data mergeabl;
one=1;    /*  set pointer to one */
merge gen_data;   /*  Dataset with item responses */
set item_par point=one;  /*  Item parameters */
set misc point=one;      /*  Model type and # of categories information */
set quad point=one;   /*  Quadrature points and priors */
rename theta1-theta&ndims = true1-true&ndims;
run;

%mend mergeabl;

%macro likehood;

/*  Calculate likelihood functions in CAT environment */

data likehood;
set mergeabl;

array model{&nitems};
array ncat{&nitems};
array ix{&nitems} ix1-ix&nitems;
array p{&nitems,&nparms} p1-p%eval(&nitems * &nparms);
array qpt{&nqpt} qpt1-qpt&nqpt;
array prior{&nqpt} prior1-prior&nqpt;
array lk{%eval(&nqpt + 1)};


* compute likelihood of response pattern at theta points - handles mixed
models;
```

```
* To do so means that the dataset weights is attached to each record;
  do i=1 to &nqpt;
        lk{i}=1;
        do j=1 to &nitems;
           resp=ix{j};
           if resp ~= &missing then
              do;
                 theta1=qpt[i];
                 if model[j]="di" then do;
                  %di;
                             end;
                         else if model[j]="gr" then do;
                              %gr;
                             end;
                       else if model[j]="pc" then do;
                              %pc;
                             end;
                         else if model[j]="gpc" then do;
                    %gpc;
                             end;
                         else if model[j]="md" then do;
                              %di;
                             end;
                 lk{i}=lk{i}*xx;
              end;
         end;
  end;
%mend likehood;

%macro eapest;
/*  Estimate ability using EAP */

data eapfull eap(keep=eap eapvar true1);
set likehood;

array model{&nitems};
array ncat{&nitems};
array ix{&nitems} ix1-ix&nitems;
array p{&nitems,&nparms} p1-p%eval(&nitems * &nparms);
array qpt{&nqpt} qpt1-qpt&nqpt;
array prior{&nqpt} prior1-prior&nqpt;
array lk{%eval(&nqpt + 1)};

/*  Now compute EAP and var{EAP} */

retain eapmean (0) eapvar (0);
zz=0;
z=0;
xx=0;
yy=0;
ssum=0;

do i=1 to &nqpt;
      xx = xx + lk{i}*qpt{i}*prior{i};
      yy = lk{i}*prior{i};
      ssum = ssum+yy;
end;

eap = xx/ssum;

do i=1 to &nqpt;
      z = z + lk{i}*prior{i}*(qpt{i} - (xx/ssum))**2;
end;

eapvar = (z/ssum);
run;

%mend eapest;

%macro infofn;
/********************************************/
/****  Item information function routine ****/
/********************************************/

/*  Computes item information for dichotomous and
    graded response items */
```

```
/*  Transpose quadrature points */
proc transpose data=quad (keep = qpt1-qpt&nqpt) out=quadt prefix=quad;
run;

data quadt;
      set quadt;
      rename quad1 = theta1;
run;

data infofull infofn (keep=theta1 info1-info&nitems);
      one=1;    /*  set pointer to one */
      merge quadt;
      set item_par point=one;
      set model_type point=one;
      set ncat_info point=one;

      array theta{&ndims} theta1-theta&ndims;     /*  For future expansion to MIRT item info */
      array model{&nitems};
      array ncat{&nitems};
      array p{&nitems,&nparms} p1-p%eval(&nitems * &nparms);

      /*  Probability functions for each response category */
      array cumprob{*} cumprob1-cumprob%eval(&maxcat);

      /*  Boundary ogives (extend +1 past # of categories) */
      array ogive{*} ogive1-ogive%eval(&maxcat+1);

      /*  Allocate space for information functions */
      array info{&nitems} info1-info&nitems;

      /*  Calculate probability functions */

      do j=1 to &nitems;
            do k=1 to &maxcat;
               cumprob[k]=.;
               end;
            do resp=0 to ncat{j}-1;
            if model[j]="di" then do;
      %di;
          end;
            else if model[j]="gr" then do;
                  %gr;
            end;

    if resp=0 then cumprob{1}=xx;
    else cumprob{resp+1}=xx+cumprob{resp};
end;

      if model[j]="di" then do;
      info[j] = (&d**2)*(p{j,1}**2)*cumprob{1}/(1-cumprob{1})*
                  ( (( (1-cumprob{1}) - p{j,3})**2)/((1 - p{j,3})**2) );
    end;

if model[j]="gr" then do;

            /*  Need to find boundary ogives, or P*
                cumprob{i} = sum (from 1 to i) of the P(category)
                  where category goes from 1 to k

                Then P*(i) = 1 - cumprob{i}
            */

              /* First assign value to lowest boundary point */
            ogive(1) = 1;


              /*  Now find boundary ogives for remaining points
                  (note that last point is always equal to zero) */

              do i=1 to ncat(j);
                   ogive(i+1)=1-cumprob(i);
              end;

            /*  Calculate information (see Baker (1991) Ch.8 p. 246)*/
      infosum=0;
```

```
                  do k=2 to ncat(j)+1;
                             infosum = infosum + ( &d*p{j,1} )**2 *
                         (  (  ogive(k-1)*(1-ogive(k-1)) - ogive(k)*(1-ogive(k)) )**2
                            /
                                        (  ogive(k-1) - ogive(k) )
                          );
                  end;
                        info[j]=infosum;
             end;

            end;

      run;

/*  Transpose dataset so that each row represents and item, columns give
    item information at each of the &nqpt quadrature points */

proc transpose data=infofn out=infofn prefix=info;
      var info1-info&nitems;
run;

data infofn;
      retain item 0;
      one=1;
      merge infofn;
      set quad point=one;
      keep info1-info&nqpt qpt1-qpt&nqpt item;
      item=item+1;
      output;
run;

%mend infofn;

%macro infomax;
/************************************************/
/**** Identify point of maximum information ****/
/************************************************/
```

```
data infomax;
retain item 0;
set item_par_full;
array x{&nparms} x1-x&nparms;

if model="di" then do;
      thetamax = x{2} + (1/(&d*x{1}))*log(0.5 + 0.5 * sqrt(1 + 8*x{3}) );
    infomax = ( (&d**2 * x{1}**2) / (8 * (1 - x{3})**2 ) ) *

               ( 1 - 20*x{3} - 8*x{3}**2 + (1 + 8*x{3})**(3/2) );
end;


item = item + 1;
output;
run;

%mend infomax;

%macro cat;
******************************************************************;
/*  SUBROUTINE CAT

      Highest-level module for executing a CAT administration

      INPUT DATASETS:   <none>
      INPUT MACRO VARIABLES:  &NSUBJ, &CATLENGTH
      OUTPUT DATASETS:  CAT
*/
******************************************************************;

      %gencat2;         /*  Simulate examinee responses to all items; CAT will select */
      %quad;                  /*  Find quadrature points and assign priors */

      %quadgrid       /*  Fine-scaled quadrature points and priors */

      /*  Prepare item sets for the CAT administration */
      /* Order of infofn and itemproc important here! */
```

```
/*
      %infofn;
*/
      %itemproc;
/*
      %inforow;
*/

      %expcon;      /*  Create row vector of exposure control parameters */

/************************************************************/
/*        BEGIN CAT ADMINISTRATION                          */
/************************************************************/

      %catloop;

      data cat;
            set catloop;
      run;

%mend cat;

%macro catloop;
*******************************************************************;
/*  SUBROUTINE CATLOOP

    Conducts a CAT administration for examinees.  Allocates
    data space for each examinee, selects items, computes provisional
    ability estimates.

    Calls the following subroutines:
            -  FINDQUAD     Finds quadrature point for current ability estimate
        -  ITEMSELECT   Selects items
            -  LIKEHOODCAT  Computes likelihood function after each
                        administered item
            -  ITEMREC      Records items exposed to examinee in ITEM_PAR_FULL

    INPUT DATASETS:  ITEM_PAR_FULL, CATLOOP
    MACRO VARIABLES:  &CATLENGTH, &nparms, &ICOUNTER,
                                        &nparms,
    OUTPUT DATASETS:  CATLOOP

*/
/**********************************************************************/

      data gen_data2;
            one=1;
            merge gen_data;
*           set inforow point=one;
            set quad point=one;
            set quadgrid point=one;
            set expcon point=one;
            set binpts point=one;
      run;

      /*  Allocate space for examinee record */
      data catloop;
            set gen_data2;

            /*  Exposure control variables */
            retain expose_s1-expose_s%eval(&bins * &nitems) 0;  /*  Selected */
            retain expose_a1-expose_a%eval(&bins * &nitems) 0;  /*  Selected and administered */
            retain seed 0;                                    /*  Seed for Sympson-Hetter random numbers */

            /*  These arrays are for items selected by the CAT */
            array admin{*} admin1-admin&catlength;  /* Item id info of administered items */
      array model{&catlength} $ model1-model&catlength; /* Item type */
      array ncat{&catlength} ncat1-ncat&catlength;  /*  Number of categories */

            array ix{*}  ix1-ix&catlength;    /* Examinee responses to items */
            array prov{*}  prov1-prov&catlength;    /* Provisional ability estimates */
      array p{&catlength,&nparms} p1-p%eval(&catlength * &nparms); /* Item parameters of selected items */
        array par{&nitems,&nparms} par1-par%eval(&nitems * &nparms);

*           array info(&nitems,&nqpt) info1-info%eval(&nitems * &nqpt); /*  Item information */
            array picked{&nitems} picked1 - picked&nitems;  /* New array for %exact */
            array expose_s{&nitems,&bins} expose_s1 - expose_s%eval(&nitems * &bins); /* Exposure: 
selected items */
            array expose_a{&nitems,&bins} expose_a1 - expose_a%eval(&nitems * &bins); /* Exposure: 
selected & administered items */
            array expcon{&nitems,&bins} expcon1 - expcon%eval(&nitems * &bins); /* Exposure control 
parameters */
            array binpt(&bins) binpt1 - binpt&bins; /*  Cut-off points for conditional exposure control */

            /*  These arrays are for the entire matrix of simulated item responses */
            array parmodel{&nitems} $ parmodel1-parmodel&nitems; /* Item type */
      array parncat{&nitems} parncat1-parncat&nitems;  /*  Number of categories */
        array parix(&nitems) parix1-parix&nitems;  /*  Simulated responses */

            /*  These arrays are for ability estimation routines */
            array qpt{&nqpt} qpt1-qpt&nqpt;                  /*  Quadrature points */
            array prior{&nqpt} prior1-prior&nqpt;            /*  Prior density */
            array lk{%eval(&nqpt + 1)};                      /*  Likelihood function */
            array post{&nqpt} post1-post&nqpt;               /*  Posterior density */
            array cumprob{*} cumprob1-cumprob%eval(&maxcat);       /* Cumulative probs for info fn */
            array ogive{*} ogive1-ogive%eval(&maxcat+1);           /* Ogives for info fn */

            array qgrid(&nqpt,&qfine) qgrid1 - qgrid%eval(&nqpt * &qfine);    /*  Fine-scaled quadrature 
grid */
            array pgrid(&nqpt,&qfine) pgrid1 - pgrid%eval(&nqpt * &qfine);    /*  Prior density for fine-
scaled grid */

            /*  Forecasting arrays */
            array forecast(&catlength) forecast1 - forecast&catlength;
            array infomeas(&catlength) infomeas1 - infomeas&catlength;

            /*  Set initial examinee ability estimate */
            thtest = 0;

            /*  Clear arrays for item selection by %exact */
            %clearexact;

            do icounter=1 to &catlength;
      /*  Find quadrature point */
      %findquad;

      /*  Select item from inforow vector */
      %itemselect;

      /*  Save item id information in admin{*} array */
      admin[icounter]=selected;

      /*  Save parameters of selected item in CATLOOP dataset */
model[icounter]=parmodel(selected);
ncat[icounter]=parncat(selected);

      do j=1 to &nparms;
            p[icounter,j]=par[selected,j];
      end;

/*  Save response to current selected item */
      ix[icounter]=parix(selected);

      /*  Estimate ability */
      %thtest;

      /*  Save ability estimate in examinee record */
      prov(icounter)=thtest;

      %if &expselect = 1 %then %do;
            /*  Experimental item selection*/
            %flip;
            %end;

      %if &expselect = 2 %then %do;
            /*  Hybrid item selection
                  Items 1-10:  Experimental
                  Items 11- :  FI  */
            if icounter <= 10 then do;
                  %flip;
```

```
                          end;

                    %end;

              %if &expselect = 3 %then %do;
                    thtest = true1;
                    %end;

              /*  Calculate test information measure for
                    relative efficiency comparisons */

              %if &infomeas = 1 %then %do; * At true theta;
                    %infotrue;
                    %end;

              %if &infomeas = 2 %then %do; * At estimated theta;
                    %infoest;
                    %end;

              /*  For variable-length CAT, check precision of measurement */
              if &cattype = 2 and thtvar <= &catprec2 then do;
                    length = icounter; /* Record length of administered test */
                    leave;
                    end;

        end;

  run;

/*  Record items that were exposed */
  %itemrec;

  /*  Trim CATLOOP dataset, keeping only relevant information */

  data catloop;
        set catloop;
        keep true1;
        keep admin1-admin&catlength;
```

```
            keep model1-model&catlength;
            keep ncat1-ncat&catlength;
            keep p1-p%eval(&catlength * &nparms);
            keep ix1-ix&catlength;
            keep prov1-prov&catlength;
            keep thtest thtvar;
            keep length; /*  Needed for variable length CATs */
            keep bin;    /*  Bin number, for exposure control */
            keep binpt1 - binpt&bins;     /*  Ability cutoff points for exposure control */
            keep forecast1 - forecast&catlength;
            keep infomeas1 - infomeas&catlength;
            keep mnc mn z;
run;

%mend catloop;

%macro itemproc;

%if (&mode = 2) or (&mode = 3) %then %do;
      %if (&expcon = 1) %then %do;
            %expconin;
            %end;
      %else %if (&expcon = 0) %then %do;
            %expconset;
            %end;
%end;

%if (&mode = 4) %then %do;
      %expconset;
      %end;

%mend itemproc;

%macro expconset;
/*  Prepares item_par_full dataset for use in CAT subroutines; initializes
      exposure control variables

      Used when an exposure control parameter file is not being loaded */
```

```
data item_par_full;
      set item_par_full;  /*  Obtain item types & parameters */
*     set infofn;         /*  Append item information       */

            /*  Initialize exposure control variables */

          retain exposed_s1 - exposed_s&bins 0;  /*  Items selected */
      retain exposed_a1 - exposed_a&bins 0;  /*  Items selected and administered */
            retain expcon1 - expcon&bins 1;        /*  Initialize exposure control parameter */

      drop qpt1-qpt&nqpt;  /*  Quad points from INFOFN not needed here */

run;

/*  Generate ability scale cutpoints for conditional Sympson-Hetter */

data binpts;
      array binpt(&bins) binpt1 - binpt&bins;
      expinc = (2 * &ub) / &bins; /* Consider range +/- &ub */
      do i = 1 to &bins;
            binpt(i) = -&ub + i*expinc;
      end;
run;

%mend expconset;

%macro expconout;

/*  Output number of bins, ability scale cutpoints, and exposure
      control parameters to a file */

data expconout;
      set item_par_full;
      file wrkdir(&expconout);
      bins = &bins;
      put itemid bins (binpt1 - binpt&bins) (7.3) (expcon1 - expcon&bins) (7.3);
run;
```

```
%mend expconout;


%macro expconin;

/*  Load first line of file to identify number of bins */

data expconin;
      infile wrkdir(&expconin) firstobs = 1 obs = 1;
      input itemid bins;
      call symput('bins',bins);
run;

/*  Now load ability scale cutpoints and exposure control parameters
      from entire file */

%let bins = %eval(&bins);

data expconin (drop = itemid);
      infile wrkdir(&expconin);
      input itemid bins binpt1 - binpt&bins expcon1 - expcon&bins;
run;

/*  Merge exposure control information with ITEM_PAR_FULL */
/*  A match merge would be preferable, but errors are currently
      encountered regarding the variable type declaration of ITEMID */

data item_par_full;
      set item_par_full;  /*  Obtain item types & parameters */
      set infofn;         /*  Append item information       */
      merge expconin;         /*  Merge exposure control parameters */

            /*  Initialize exposure control variables */

          retain exposed_s1 - exposed_s&bins 0;  /*  Items selected */
      retain exposed_a1 - exposed_a&bins 0;  /*  Items selected and administered */
      drop qpt1-qpt&nqpt;  /*  Quad points from INFOFN not needed here */

run;

/*  Load ability scale cutpoints for conditional Sympson-Hetter */

data binpts (keep = binpt1 - binpt&bins);
      set item_par_full;
run;

%mend expconin;

%macro gencat2;

/*  Generate full pattern of item responses before CAT executes.  This
      method saves computational time in SAS */

%seeds;
%thetagen;
%itempar;

%let nitems=&npool;

%cattype; /*  Set program parameters for fixed- or variable-length CAT */
%mergegen;
%gendata;

      /*  Rename item type & parameter information for use with CATLOOP */

      data gen_data;
            set gen_data;
            rename p1-p%eval(&nitems * &nparms) = par1-par%eval(&nitems * &nparms);
            rename model1-model&nitems = parmodel1-parmodel&nitems;
            rename ncat1-ncat&nitems = parncat1-parncat&nitems;
            rename ix1-ix&nitems = parix1 - parix&nitems;
            rename theta1-theta&ndims = true1-true&ndims;

      run;
```

```
%mend gencat2;

%macro cattype;
%global catprec2;
/*  Set parameters to conform to fixed- or variable-length CAT */

/*  Here, macro variables declared in the main program may be reset */

/*  Note that this routine does not affect data sets, instead it
    sets values for macro variables */

/*  INPUT VARIABLES:  &CATTYPE, &CATLENGTH, &CATPREC, &NITEMS
      OUTPUT VARIABLES:  &CATPREC2    */

      /*  Insure that length limits for variable-length CAT do not exceed
    item pool size */

      %if &catlimit>&nitems %then %do;
            %let catlimit=&nitems;
            %end;

      /*  Fixed-length CAT parameters */
      %if &cattype=1 %then %do;
            %let catprec = 0;
            %end;

      /*  Variable-length CAT parameters */
      %else %if &cattype=2 %then %do;
            %let catlength=&catlimit;
            %end;

      %let catprec2=%sysevalf(&catprec**2); /* Square of standard error of measurement */

%mend cattype;

%macro itemselect;
*****************************************************;
```

```
/*  SUBROUTINE ITEMSELECT

    Selects items in CAT environment, and updates
    exposure information

      Modified to execute WITHIN data step; increases
    efficiency
*/
*****************************************************;

      /*  Use Sympson-Hetter for exposure control */
      /*  When flag is equal to one, an item is selected */
      /*  If exposure control is not being used, flag is always equal to one */

      if &expcon=1 then do;   /* &expcon = 1 <==> exposure control ON */
            flag = 0;
            end;
      else do;                /* &expcon = 0 <==> exposure control OFF */
            flag = 1;
            end;

      do loop = 1 to &nitems;  /* More robust than a do while or until loop */

                  if &itemsel=1 then do;  /*Max info selection */
                        %exact;
                        end;

                  if &itemsel=2 then do;
                        %fii;
                        end;

                  /*  NOTE:  This code has been modified to conform with
                        macro %exact */

                  /*  Item has now been selected, and pointer to item is
                        variable "selected" */

                  /*  Find proper bin, based on examinee ability, to record
```

```
                        item exposure information */

                  do i = &bins to 1 by -1; /* Descending search */
                        if true1 <= binpt(i) then do;
                              bin = i;
                              end;
                        else if true1 > binpt(&bins) then do;
                              bin = &bins;
                              end;
                  end;

                  /*  Increase "selected" exposure counter for item */
                  expose_s(selected,bin)= expose_s(selected,bin)+1;

            /*  Apply Sympson-Hetter technique */
                  if &expcon = 1 then do;
                        call ranuni(seed,r);
                        if r <= expcon(selected,bin) then do;  /*  Administer item */
                              expose_a(selected,bin) = expose_a(selected,bin) + 1;
                              flag = 1;
                        end;
                  end;

                  if flag = 1 then leave; /*  Leave loop when item is selected for administration */

      end;


%mend itemselect;

%macro maxinfo;
/*  Maximum information item selection */

                  maxvalue=0;
                  selected=0;

                  do search=1 to &nitems;
                        if info(search,qpointer)>maxvalue then do;
```

```
                              maxvalue=info(search,qpointer);
                              selected=search;
                              end;
                  end;

%mend maxinfo;

%macro fip;
/*  Fisher information with posterior weight function */

      maxvalue=0;
      selected=0;

      do search=1 to &nitems;
          integral = 0;
            do i = 1 to &nqpt;
                  integral = integral + post{i}*info(search,i);
            end;
        if integral > maxvalue then do;
                  maxvalue = integral;
                  selected = search;
            end;
    end;

%mend fip;

%macro clearexact;

do i = 1 to &nitems;
      picked(i) = 0;
end;

%mend clearexact;

%macro exact;
/*  Maximum information item selection without
    tabled information values; computes information
      at thtest each time */
```

```
maxvalue = 0;
selected = 0;

do search = 1 to &nitems;

      /*  An item has already been administered
            if picked(*) = 1; skip this item
            if already administered */

      if picked(search) ~= 1 then do;

      /*  Compute information at thtest for item */

                  j = search;  * assign index variable;

            do k=1 to &maxcat;
               cumprob[k]=.;
               end;

            do resp=0 to parncat{j}-1;
                        /*  Modified %di macro to reflect change from
                              p(*) to par(*) array */
                        xx=.;
                  xx=par{j,3}+(1-par{j,3})/(1+exp(-&d*par{j,1}*(thtest-par{j,2})));
                        if resp=0 then xx=1-xx;
                        /*  End modified %di macro */

                if resp=0 then cumprob{1}=xx;
                else cumprob{resp+1}=xx+cumprob{resp};
            end;

            iteminfo = (&d**2)*(par{j,1}**2)*cumprob{1}/(1-cumprob{1}) *
                              ( (( (1-cumprob{1}) - par{j,3})**2)/((1 - par{j,3})**2) );

      /*  Compare information & update */

                  if iteminfo > maxvalue then do;
```

```
                        maxvalue = iteminfo;
                        selected = search;
                        end;

            end;

end;

/*  Record item that has been picked */

picked(selected) = 1;

%mend exact;

%macro fii;
/*  Item selection by Fisher interval information (FII),
    uses exact solution to integral for 3P case */

*     if icounter > 1 then do;   /*  An interval exists only when icounter > 1 */

            /*  Reset search variables */
            maxvalue = 0;
            selected = 0;

            /*  Form interval as suggested by Chen, Ankenmann, & Chang (APM, 2000), p. 248 */
            lbf = thtest - &zfii * 1/(sqrt(icounter) + 1);
            ubf = thtest + &zfii * 1/(sqrt(icounter) + 1);

            /*  Insure interval is bounded by [-&ub, &ub] */
            if lbf < -&ub then do;
                  lbf = -&ub;
                  end;

            if ubf > &ub then do;
                  ubf = &ub;
                  end;

            /*  Search through items for maximum FII */
```

```
do search = 1 to &nitems;

      /*  An item has already been administered
            if picked(*) = 1; skip this item
            if already administered */

      if picked(search) ~= 1 then do;

            /*  Compute terms for exact solution to 3P FII */
            A = &d * par(search,1);
            B = par{search,2};
            C = par{search,3};
            x1 = exp(A*B);
            x2 = exp(A*lbf);
            x3 = exp(A*(lbf - B));
            x4 = exp(A*ubf);
            x5 = exp(A*(ubf - B));
            fii = A * ( (C-1)*x1 + C*(x1 + x2)*log(1 + x3) - C*(x1 + x2)*log(C + x3) )
                          /
                    ( (C-1)*(x1 + x2) )

                    -

                    A * ( (C-1)*x1 + C*(x1 + x4)*log(1 + x5) - C*(x1 + x4)*log(C + x5) )
                        /
                    ( (C-1)*(x1 + x4) );

            /*  Compare FII index & update */

            if fii > maxvalue then do;
                  maxvalue = fii;
                  selected = search;
                  end;

      end; /* End conditional on picked(search) */

end;  /* Iterative loop for <search> */
```

```
            /*  Record item that has been picked */
            picked(selected) = 1;

*       end; /* Conditional on <icounter> */

*       else do;    /*  Otherwise, perform max info (FI) item selection */
*           %exact;
*           end;


            /*  NOTES

            Chen et al. (APM, 2000) suggest earlier in their paper that
            a confidence interval be formed based on the information measure
            at the current provisional ability estimate (see pp. 243 - 244).
            However, when this confidence interval is used, FII item selection
            performs quite poorly.

            The following code is provided in order to test out such an
            interval. */

            /*  Find information at current ability estimate thtest */
*           %infoest2;

            /*  Find lower and upper bounds for interval */
*           lbf = thtest - &zfii * (1/sqrt(testinfo));
*           ubf = thtest + &zfii * (1/sqrt(testinfo));

%mend fii;

%macro fii2;

/*  FII by approximation to integral */

    if icounter > 1 then do;   /*  An interval exists only when icounter > 1 */

            /*  Reset search variables */
            maxvalue = 0;
```

```
selected = 0;

/*  Find information at current ability estimate thtest */
%infoest2;

/*  Form interval as suggested by Chen, Ankenmann, & Chang (APM, 2000), p. 248 */
lbf = thtest - &zfii * 1/(sqrt(icounter) + 1);
ubf = thtest + &zfii * 1/(sqrt(icounter) + 1);

/*  Insure interval is bounded by [-&ub, &ub] */
if lbf < -&ub then do;
      lbf = -&ub;
      end;

if ubf > &ub then do;
      ubf = &ub;
      end;

increment = (ubf - lbf)/20;

/*  Search through items for maximum FII */
do search = 1 to &nitems;

      /*  An item has already been administered
            if picked(*) = 1; skip this item
            if already administered */

      if picked(search) ~= 1 then do;

            fii = 0;
            j = search;

            do thtpt = lbf to ubf by increment;

                  /*  Get info */

            do k=1 to &maxcat;
            cumprob[k]=.;
```

```
                  end;

                  do resp=0 to parncat{j}-1;
                              /*  Modified %di macro to reflect change from
                              p(*) to par(*) array */
                              xx=.;
                        xx=par{j,3}+(1-par{j,3})/(1+exp(-&d*par{j,1}*(thtpt-par{j,2})));
                              if resp=0 then xx=1-xx;
                              /*  End modified %di macro */

            if resp=0 then cumprob{1}=xx;
            else cumprob{resp+1}=xx+cumprob{resp};
                  end;

                  iteminfo = (&d**2)*(par{j,1}**2)*cumprob{1}/(1-cumprob{1}) *
                        ( (( (1-cumprob{1}) - par{j,3})**2)/((1 - par{j,3})**2) );

                        fii = fii + iteminfo;

                  end;

                  /*  Compare FII index & update */

                  if fii > maxvalue then do;
                        maxvalue = fii;
                        selected = search;
                        end;

            end; /* End conditional on picked(search) */

      end;  /* Iterative loop for <search> */

      /*  Record item that has been picked */
      picked(selected) = 1;

end; /* Conditional on <icounter> */

else do;    /*  Otherwise, perform max info (FI) item selection */
```

```
            %exact;
            end;

%mend fii2;



%macro findquad;

      qpointer= int( (thtest + &ub) / &xinc ) + 1;
      if qpointer < 1 then qpointer = 1;
            else if qpointer > &nqpt then qpointer = &nqpt;

%mend findquad;

%macro inforow;
/*  Convert the item information matrix in INFOFN to a row for
appending to CATLOOP dataset */

      data inforow;
            set infofn;
            array y{*} info1-info&nqpt;
            keep p;
            do j=1 to &nqpt;
            p=y{j};
            output;
            end;
      run;

/*  Create a row vector with item parameters as elements */
      proc transpose data=inforow out=inforow prefix=info;
            var p;
      run;

%mend inforow;

%macro likehoodcat;
```

```
/*  Similar to subroutine LIKEHOOD except this is a segment of
    SAS statements, rather than a self-contained DATA step */

  do i=1 to &nqpt;
            theta1=qpt[i]; /* Pass quad point as theta value */
        lk{i}=1;
            do j=1 to icounter;
           resp=ix{j};
           if resp ~= &missing then
              do;
                 if model[j]="di" then do;
                  %di;
                              end;
                         else if model[j]="gr" then do;
                              %gr;
                              end;
                     lk{i}=lk{i}*xx;
              end;
        end;
  end;
%mend likehoodcat;

%macro thtest;
/*  Estimate ability using particular estimation
      method (ML, MAP, or EAP)

      Newton - Raphson:
      ML if &est=1; MAP if &est=2;

      EAP if &est=3;

      Grid search:
    ML if &est=4; MAP if &est=5 */

                  if &est=3 then do;
                        %eapestcat;
                        thtest=eap;
                        thtvar=eapvar;
                        end;
                  else if &est=4 then do;  /* Grid search for MLE */
                        %maxsearch;
                        thtest=maxpoint;
                        %modalse; /* Find asymptotic SE of ML estimator */
                        thtvar=mlvar;
                        end;
                  else if &est=5 then do;  /* Grid search for MAP */
                        %maxsearch;
                        thtest=maxpoint;
                        %modalse;
                        thtvar=mapvar;
                        end;
                  else if &est=1 then do; /* Newton-Raphson for MLE */
                        %newton;
                        if callnewt = 1 then do; /* Callnewt = 1 for non-perfect response pattern */
                              %modalse;
                              thtvar=mlvar;
                              end;
                        else do;  /*  Perfect response pattern; ability assigned to +/- &ub, no SE
available */
                              thtvar = 0;
                              end;
                        end;
                  else if &est=2 then do; /* Newton-Raphson for MAP */
                        %newton;
                        %modalse;
                        thtvar=mapvar;
                        end;
                  else if &est=6 then do; /* Golden search strategy (GSS) */
                        %gss;
                        %modalse;
                        thtvar=mlvar;
                        end;
%mend thtest;


%macro post;
```

```
/*  Compute posterior distribution, store in post{} array */

      do i=1 to &nqpt;
            post[i] = lk[i]*prior[i];
      end;

%mend post;

%macro eapestcat;
/*  Similar to subroutine EAPEST except this is a segment of
    SAS statements */

      /*  Estimate ability using EAP */

      /*  Compute likelihood function */
      %likehoodcat;

      /*  Calculate posterior distribution */
      %post;

      zz=0;
      z=0;
      xx=0;
      yy=0;
      ssum=0;

      do i=1 to &nqpt;
            xx = xx + post{i}*qpt{i};
            yy = post{i};
            ssum = ssum+yy;
      end;

      eap = xx/ssum;

      do i=1 to &nqpt;
            z = z + post{i}*(qpt{i} - (eap))**2;
      end;
      eapvar = (z/ssum);

%mend eapestcat;

%macro maxsearch;
/*  Experimental macro for global maximum search */

/*  Rough-scaled search for maximum */

      /*  Compute likelihood function */
      %likehoodcat;

      /*  Compute posterior distribution*/
      %post;

      /*  Initialize maxvalue & maxpoint */
      maxvalue = 0;
      maxpoint = 0;

      do i=1 to &nqpt;
            if post{i} > maxvalue then do;
                  maxvalue = post{i};
                  maxpoint = i;
                  end;
      end;


      /*  Re-initialize maxvalue & create maxpoint1, maxpoint2 */
      maxvalue = 0;
      maxpoint1 = 0;
      maxpoint2 = 0;

      /*  Perform fine-scaled search
      Search will be +/- 1 rough-scaled quadrature point from RPOINT */

      /*  Identify lower and upper bounds for search */
      lb = max(1,maxpoint-1);
      ub = min(&nqpt,maxpoint);
      /*  Calculate likelihoods on finer scale */
      do i=lb to ub;
            do k=1 to &qfine;
                  theta1=qgrid[i,k]; /* Pass quad point as theta value */
            like=1*pgrid[i,k];
                  do j=1 to icounter;
                  resp=ix{j};
                  if resp ~= &missing then
                  do;
                  if model[j]="di" then do;
                        %di;
                                    end;
                              else if model[j]="gr" then do;
                                    %gr;
                                    end;
                              like=like*xx;
                         end;
             end;
                   if like > maxvalue then do;
                       maxvalue = like;
                         maxpoint1 = i;
                         maxpoint2 = k;
                         end;
           end;
      end;

      maxpoint = qgrid(maxpoint1,maxpoint2);

%mend maxsearch;

%macro modalse;
/*  Use information function at the ML estimate to compute
    asymptotic standard error of the estimate */

/*  The current test information I(theta) may be computed by
    summing the individual item information functions I_i(theta)
    where theta is equal to the ML estimate.
    Then SE(theta_ML) = 1 / sqrt(I(theta_ML)) */

/*  Routine makes use of cumprob{*}, ogive{*}, ncat{*},
    model{*}, and p{*} arrays */

/*  The following code is taken from the %infofn routine
    The info{*} array is not used; item information is stored in
    the variable iteminfo; test info in variable testinfo */

      /*  Info for each item is found at theta = theta_ML.  This variable
        is identified as thtest.  The %di and %gr routines are functions
        of the variable theta1 */

      theta1 = thtest;
      testinfo = 0;

      do j=1 to icounter; /*j index insures compatibility with %di and %gr routines*/
            do k=1 to &maxcat;
               cumprob[k]=.;
               end;
            do resp=0 to ncat{j}-1;
                        if model[j]="di" then do;
                  %di;
                      end;
                        else if model[j]="gr" then do;
                              %gr;
                        end;

                if resp=0 then cumprob{1}=xx;
                else cumprob{resp+1}=xx+cumprob{resp};
            end;

                  if model[j]="di" then do;
                  iteminfo = (&d**2)*(p{j,1}**2)*cumprob{1}/(1-cumprob{1})*
                              ( (( (1-cumprob{1}) - p{j,3})**2)/((1 - p{j,3})**2) );
                end;
    if model[j]="gr" then do;

                /*  Need to find boundary ogives, or P*
                    cumprob{i} = sum (from 1 to i) of the P(category)
                      where category goes from 1 to k

                    Then P*(i) = 1 - cumprob{i}
                */

                  /* First assign value to lowest boundary point */
                ogive(1) = 1;


                  /*  Now find boundary ogives for remaining points
                      (note that last point is always equal to zero) */

                  do i=1 to ncat(j);
                       ogive(i+1)=1-cumprob(i);
                  end;

                /*  Calculate information (see Baker (1991) Ch.8 p. 246)*/
          infosum=0;
          do k=2 to ncat(j)+1;
                     infosum = infosum + ( &d*p{j,1} )**2 *
                 (  (  ogive(k-1)*(1-ogive(k-1)) - ogive(k)*(1-ogive(k)) )**2
                    /
                                ( ogive(k-1) - ogive(k) )
                  );
          end;
                iteminfo=infosum;
     end;

           testinfo = testinfo + iteminfo;

    end;

mlvar = 1/testinfo;
```

```
            /*  If a normal prior N(0,1) is used, then it can be shown that
                I(theta_MAP) = I(theta_ML) + 1

                More generally, for a normal prior with variance = sigma**2,
                I(theta_MAP) = I(theta_ML) + (1/sigma**2) */

            mapvar = 1 / (testinfo + (1/&sigma2) );

%mend modalse;


%macro infotrue;

/*  Calculate test information at true theta value */
      theta1 = true1;
      testinfo = 0;

      do j=1 to icounter; /*j index insures compatibility with %di and %gr routines*/
            do k=1 to &maxcat;
               cumprob[k]=.;
               end;
            do resp=0 to ncat{j}-1;
                        if model[j]="di" then do;
                  %di;
                            end;

                if resp=0 then cumprob{1}=xx;
                else cumprob{resp+1}=xx+cumprob{resp};
            end;

                  if model[j]="di" then do;
                  iteminfo = (&d**2)*(p{j,1}**2)*cumprob{1}/(1-cumprob{1}) *
                              ( (( (1-cumprob{1}) - p{j,3})**2)/((1 - p{j,3})**2) );
                  end;

            lastp = 1 - cumprob{1};  /*  Save prob fn from last administered item */
                  testinfo = testinfo + iteminfo;
      end;
      /*  Save test information in array */

      infomeas(icounter) = testinfo;

%mend infotrue;

%macro infoest;

/*  Calculate test information at estimated theta value (provisional estimate) */
      theta1 = prov(icounter);
      testinfo = 0;

      do j=1 to icounter; /*j index insures compatibility with %di and %gr routines*/
            do k=1 to &maxcat;
               cumprob[k]=.;
               end;
            do resp=0 to ncat{j}-1;
                        if model[j]="di" then do;
                  %di;
                            end;

                if resp=0 then cumprob{1}=xx;
                else cumprob{resp+1}=xx+cumprob{resp};
            end;

                  if model[j]="di" then do;
                  iteminfo = (&d**2)*(p{j,1}**2)*cumprob{1}/(1-cumprob{1}) *
                              ( (( (1-cumprob{1}) - p{j,3})**2)/((1 - p{j,3})**2) );
                  end;

            lastp = 1 - cumprob{1};  /*  Save prob fn from last administered item */
                  testinfo = testinfo + iteminfo;
      end;

      /*  Save test information in array */

      infomeas(icounter) = testinfo;
```

```
%mend infoest;

%macro infoest2;

      /*  This routine is called before the item indicated by <icounter>
        is administered */

      theta1 = thtest;
      testinfo = 0;

      do j=1 to (icounter-1); /*j index insures compatibility with %di and %gr routines*/
            do k=1 to &maxcat;
               cumprob[k]=.;
               end;
            do resp=0 to ncat{j}-1;
                        if model[j]="di" then do;
                  %di;
                            end;

                if resp=0 then cumprob{1}=xx;
                else cumprob{resp+1}=xx+cumprob{resp};
            end;

                  if model[j]="di" then do;
                  iteminfo = (&d**2)*(p{j,1}**2)*cumprob{1}/(1-cumprob{1}) *
                              ( (( (1-cumprob{1}) - p{j,3})**2)/((1 - p{j,3})**2) );
                  end;

                  testinfo = testinfo + iteminfo;
      end;

%mend infoest2;

%macro flip;

      ztol = &ztol;  /*  Set z-tolerance, in S.D. units */
      /*  Find observed proportions */
            mn = 0;
            do i = 1 to icounter;
                  mn = mn + ix(i);
            end;

            mn = mn / icounter;

      /*  Find expected proportions, given items (adjusts for c parameters) */
            mnc = 0;
            varc = 0;

            do i = 1 to icounter;
                pc = p{i,3} + (1 - p{i,3}) /
                                 (1 + (2 / (1 + sqrt(1 + 8*p{i,3}) ) ) );
                  mnc = mnc + pc;

                  varc = varc + pc*(1-pc);   /* Variance of sum is sum of variance for independent
                                                observations */
            end;

            mnc = mnc / icounter;

            varc = varc / (icounter**2);

*           mn = observed;
*           mnc = expected;

      /*  Compute standard error for these proportions */
            se = sqrt( varc );

      /*  Compute z-statistic */
            z = (mn - mnc) / se;

            /*  Conduct hypothesis tests */

            flag = 1;
            if z < -ztol then do;  /*  Reject null on negative side */
                  newp = mnc - ztol*se;
                  end;

            else if z > ztol then do;  /*  Reject null on positive side */
                  newp = mnc + ztol*se;
                  end;

            else do;   /*  Do not reject null */
                  flag = 0;
                  end;

            /*  Null hypothesis rejected when flag = 1 */
            if flag = 1 then do;

                  /*  If flag=1, alternative procedure is used.

                        Use average ICC to locate new ability estimate */

                  %alt;

                  /*  Assign to thtest */

                  thtest = forecast(icounter);

            end;

%mend flip;

%macro alt;

/*  Alternative procedure under ML estimation */

      resp = 1;  * For %di routine;
      lb = -&ub;
      ub = &ub;

      /*  Find inverse of newp through the
```

```
        average ICC; inverse is the new theta
        estimate */

/*      Use method of halving */

  do i = 1 to 15;
        midpoint = (lb+ub)/2;

        /*  Get lower bound prob */
        lbp = 0;
        theta1 = lb;
        do j = 1 to icounter;
              %di;
              lbp = lbp + xx;
        end;
        lbp = lbp / icounter;  *Take average;

        /*  Get midpoint prob */
        midp = 0;
        theta1 = midpoint;
        do j = 1 to icounter;
              %di;
              midp = midp + xx;
        end;
        midp = midp / icounter; *Take average;

        /*  Get upper bound prob */
        ubp = 0;
        theta1 = ub;
        do j = 1 to icounter;
              %di;
              ubp = ubp + xx;
        end;
        ubp = ubp / icounter; *Take average;

        /*  Check sections & update */
        if  newp < midp then do;
              ub = midpoint;
```

```
                  end;
            else do;
                  lb = midpoint;
                  end;

      end;

            forecast(icounter) = midpoint;

%mend alt;

%macro newton;
/*  Uses Newton-Raphson procedure to find ML and MAP ability estimates */

/*  Note that when the test information function is less than 1, this procedure
      calls %maxsearch to perform a grid search */

      callnewt = 1;  /*  Set flag initially to call Newton-Raphson */

      /*  Tabulate number correct score */
      correct=0;

      do i=1 to icounter;
            if ix[i]=1 then correct=correct+1;
      end;

      /*  Check for perfect response patterns before employing Newton-Raphson on
            MLE */

      if &est = 1 then do;  /* &est = 1 is MLE */

            if correct = 0 then do;  /* Response pattern all incorrect */
                  thtout = -&ub;
                  callnewt = 0; /* Flag to pass to Newton-Raphson, do not execute */
                  end;
            else if correct = icounter then do; /* Response pattern all correct */
                  thtout = &ub;
                  callnewt = 0; /* Flag to pass to Newton-Raphson, do not execute */
            end;

      end;


/** Newton-Raphson procedure **/
/*  Iterates using the following theta values:
            thtin:  input theta value
            thtout:  output theta value */

if callnewt = 1 then do;

      /*  Set starting value to average of b-parameters */
            stval = 0;
            do i = 1 to icounter;
                  stval = stval + p{i,2};
            end;
            thtin = stval / icounter;

      /*  Reset iteration counter */
            iters = 0;

      /*  Minimum test information required for stability */
            mininfo = 1*&d**2;

      /*  Begin loop */

      do until (iters=&newton_max or thtdiff < &crit);

            iters = iters + 1;

            /*  Obtain first and second derivatives of log likelihood at thtin */
            %deriv;

            /*  If testinfo is too small, perform grid search instead and
                do not perform Newton step */
            if testinfo < mininfo then do;
                  %maxsearch;
```

```
                        thtout = maxpoint;
                        leave; /* Leave do-until loop */
                        end;

                  /*  Otherwise, perform Newton-Raphson step */

                        thtout = thtin - (derivsum / (-1*testinfo) );
                        thtdiff = abs( thtin - thtout ); /* Find abs diff between estimates */
                        thtin = thtout; /*  Feed the thtout value as thtin for next iteration */



            end; /*do-until loop */

      end; /* conditional: callnewt */

      thtest = thtout;  /* Final theta estimate (thtest) */

%mend newton;

%macro deriv;
/*  Calculate first and second derivatives for Newton-Raphson */

            /*  Compute item and test information at current ability estimate (thtin) */
            /*  Fisher scoring used for second derivative, first derivative from
                  this relationship */

            theta1 = thtin;
            testinfo = 0;     /* Fisher scoring for second derivative */
            derivsum = 0;     /* Sum of first derivatives */
            mu = &mean;       /* Mean of prior dist_n */
            sigma2 = &sigma2; /* Variance of prior dist_n */

            do j=1 to icounter;  /*j index insures compatibility with %di and %gr routines*/
            do k=1 to &maxcat;
               cumprob[k]=.;
               end;
            do resp=0 to ncat{j}-1;
            if model[j]="di" then do;
      %di;
                end;

    if resp=0 then cumprob{1}=xx;
    else cumprob{resp+1}=xx+cumprob{resp};
end;

      if model[j]="di" then do;
      iteminfo = (&d**2)*(p{j,1}**2)*cumprob{1}/(1-cumprob{1})*
                  ( (( (1-cumprob{1}) - p{j,3})**2)/((1 - p{j,3})**2) );
      end;

      /*  First derivative for dichotomous items */
            if model[j]="di" then do;
                  /* dPdt = (a/norm(a)) * sqrt(I * P * Q) */
            /* dlnLdt = [u - P / P(1-P)] * dPdt */
                  /* Simplification of dlnLdt follows */
                  dlnLdt =  ( ix[j] - (1 - cumprob{1}) ) * ( p{j,1}/abs(p{j,1}) ) *
                          sqrt( iteminfo / ( (1 - cumprob{1}) * cumprob{1} ) );

                  /* Could also use direct derivative: */
                  /*dlnLdt = p{j,1}*(ix[j]-(1-cumprob{1}))*((1-cumprob{1})-p{j,3}) /
                    ((1-cumprob{1})*(1-p{j,3}));
                  */
                  end;

      /*  First derivative for graded response items */
            /*  [Code needs to be written here]*/

      testinfo = testinfo + iteminfo;
      derivsum = derivsum + dlnLdt;

end;

/*  If using MAP, must take into account info from prior */
if &est = 2 then do;
            /*  For N(mu,sigma**2) prior density p = p(theta)
```

```
                              dlnp/dt = - (theta - mu) / sigma**2
                              d2lnp/dt2 = - (1 / sigma**2)
                              I = -E{d2lnp/dt2}
                        */

                        derivsum = derivsum - ( (theta1 - mu) / sigma2);
                        testinfo = testinfo + (1 / sigma2);
                        end;


%mend deriv;

%macro gss;
/*  Ability estimation by golden search strategy (GSS) */

/*  Uses algorithm as described by Xiao(1999, APM) */

      /*  Reset flag, when flag = 1 an ability estimate has been found */
      flag = 0;

      /*  Specify lower and upper bounds for initial search interval */
      lb = -&ub;
      ub = &ub;

      do gss = 1 to 20;  /*  Outermost loop for GSS */

            /*  Find midpoint of current search interval */
            midpoint = (lb + ub) / 2;

            /*  Now find optimally-weighted observed and expected scores */

                  obsscore = 0;
                  expscore = 0;
                  expvar = 0;
                  minscore = 0;

                  theta1 = midpoint;
                  resp = 1;
```

```
do j = 1 to icounter;

      /*  Find P(theta1 = midpoint)
        This is returned as variable <xx> */
      %di;

      /*  Compute optimal scoring weight */
      weight = &d*p{j,1}*(xx - p{j,3})
                /
                    ( (1 - p{j,3}) * xx );

      expscore = expscore + (weight * xx);

      expvar = expvar + (weight**2 * xx * (1-xx) );

      obsscore = obsscore + (weight * ix(j));

      minscore = minscore + (weight * p{j,3});

end;

/*  Check that observed score is not smaller than
    minimum score; if it is, set observed score
    to minimum score */

if obsscore < minscore then do;
      obsscore = minscore;
      end;

/*  Conduct hypothesis test */
z = (obsscore - expscore) / sqrt(expvar);

/*  Xiao recommends abs(z_critical) = 0.7 */

if z < -0.7 then do;
      /*  Adjust upper bound of interval by golden ratio */
      ub = lb + ((sqrt(5) - 1)/2)*(ub - lb);
```

```
                        end;

                  else if z > 0.7 then do;
                        /*  Adjust lower bound of interval by golden ratio*/
                        lb = ub - ((sqrt(5) - 1)/2)*(ub - lb);
                        end;

                  else do;
                        /*  Leave loop */
                        flag = 1;
                        end;

                  if flag = 1 then leave;

      end;  /*  Outermost do-loop */

      /*  Assign theta1 to thtest */
      thtest = theta1;

%mend gss;

%macro itemrec;
/*  This macro updates the item pool with exposure information
    from the CAT administration.

      INPUT DATASET:  CATLOOP, ITEM_PAR_FULL
      OUTPUT DATASET:  EXPOSURE, ITEM_PAR_FULL

*/

/*  Experimental  ....... attempt to make this segment more efficient */
      /*  Extract exposure information from catloop dataset first */
      data exposure_info;
            set catloop;
            keep expose_a1-expose_a%eval(&nitems * &bins);
            keep expose_s1-expose_s%eval(&nitems * &bins);
            keep binpt1 - binpt&bins;
            keep bin;
```

```
run;

/*  Obtain final tally of selected & administered items */
data exposure_a;
      set exposure_info;
      /*set catloop;*/
      keep expose_a1-expose_a%eval(&nitems * &bins);
      if _n_=&nsubj; /* Save last observation, has final tally */
run;

/*  Obtain final tally of selected items */
data exposure_s;
      /* set catloop; */
      set exposure_info;
      keep expose_s1-expose_s%eval(&nitems * &bins);
      if _n_=&nsubj; /* Save last observation, has final tally */
run;

/*  Compute tally of number of subjects within each ability
      group bin, also record ability cutoff points */

data exposure_n;
      /* set catloop; */
      set exposure_info;
      keep binsize1 - binsize&bins;
      keep binpt1 - binpt&bins;
      array binsize(&bins) binsize1 - binsize&bins;
      array binpt(&bins) binpt1 - binpt&bins;
      retain binsize1 - binsize&bins 0;

      /*  Variable bin is a part of CATLOOP dataset */
      binsize(bin) = binsize(bin) + 1;

run;

/*  Obtain final tally of number of subjects within each bin */
data exposure_n;
      set exposure_n;
      array binsize(&bins) binsize1 - binsize&bins;
      keep binsize1 - binsize&bins;
      keep binpt1 - binpt&bins;
      if _n_ = &nsubj;

      /*  Must insure that binsize does not equal zero; this quantity
            will appear in a denominator in macro SYMPHET */
      do i = 1 to &bins;
            if binsize(i) = 0 then do;
                  binsize(i) = 1;
                  end;
      end;

run;

/*  Selected & administered items */
/*  Convert item exposure information, currently a row vector, into
      a dataset with &nitems rows and &bins columns */

data exposure_a2 (keep = expose_ac1 - expose_ac&bins);
      set exposure_a;
      array expose_a(&nitems,&bins) expose_a1 - expose_a%eval(&nitems * &bins);
      array expose_ac(&bins) expose_ac1 - expose_ac&bins;
      do i=1 to &nitems;
            do j=1 to &bins;
                  expose_ac(j) = expose_a(i,j);
            end;
            output;
      end;
run;

/*  Selected (but not necessarily administered) items */
/*  Convert item exposure information, currently a row vector, into
      a dataset with &nitems rows and &bins columns */

data exposure_s2 (keep = expose_sc1 - expose_sc&bins);
      set exposure_s;
      array expose_s(&nitems,&bins) expose_s1 - expose_s%eval(&nitems * &bins);
```

```
            array expose_sc(&bins) expose_sc1 - expose_sc&bins;
            do i=1 to &nitems;
                  do j=1 to &bins;
                        expose_sc(j) = expose_s(i,j);
                  end;
                  output;
            end;
      run;

      /*  Increase the exposure counters in ITEM_PAR_FULL for each item */

      data item_par_full /* drop some vars later */;
            one = 1;
            set item_par_full;
            set exposure_n point=one;
          merge exposure_a2 exposure_s2;
            array expose_ac(&bins) expose_ac1 - expose_ac&bins; /* From CATLOOP */
            array expose_sc(&bins) expose_sc1 - expose_sc&bins; /* From CATLOOP */
            array exposed_a(&bins) exposed_a1 - exposed_a&bins; /* From ITEM_PAR_FULL */
            array exposed_s(&bins) exposed_s1 - exposed_s&bins; /* From ITEM_PAR_FULL */

            do i = 1 to &bins;
                  exposed_a(i) = exposed_a(i) + expose_ac(i);
                  exposed_s(i) = exposed_s(i) + expose_sc(i);
            end;
      run;

%mend itemrec;

%macro setbinpts;
/*  Generate ability scale cutpoints for conditional Sympson-Hetter */

data binpts;
      array binpt(&bins) binpt1 - binpt&bins;
      expinc = (2 * &ub) / &bins; /* Consider range +/- &ub */
      do i = 1 to &bins;
            binpt(i) = -&ub + i*expinc;
      end;
```

```
run;

%mend setbinpts;

%macro symphettype;
/*  Determine if Sympson-Hetter technique will be conditional
      or unconditional */

      %if (&cond = 1) %then %do;
            /*  Create a nearly uniform distribution of ability.  Use the
                  transformation procedure in %thetagen by setting transformation
                  parameters for a platykurtic distribution of kurtosis = -1 */
          %let transb = 1.2210;
            %let transc = 0;
            %let transd = -0.0802;
            %end;

      %if (&cond = 0) %then %do;
            %let bins = 1;
            %end;

%mend symphettype;

%macro symphet;
/*  Updates the exposure control parameters according to the
    Sympson & Hetter technique */

      data item_par_full;
            target = &target;
            set item_par_full;
            array prob_s(&bins) prob_s1 - prob_s&bins;
            array prob_a(&bins) prob_a1 - prob_a&bins;
            array exposed_s(&bins) exposed_s1 - exposed_s&bins;
            array exposed_a(&bins) exposed_a1 - exposed_a&bins;
            array expcon(&bins) expcon1 - expcon&bins;
            array binsize(&bins) binsize1 - binsize&bins;

            do i = 1 to &bins;
```

```
                  prob_s(i) = (exposed_s(i) / binsize(i));
                  prob_a(i) = (exposed_a(i) / binsize(i));
                  if prob_s(i) > target then do;
                        expcon(i) = target / prob_s(i);
                        end;
                  else do;
                        expcon(i) = 1;
                        end;
            end;

      run;


%mend symphet;

%macro comparex;
/*  Compares observed exposure rates with target exposure rates, and sets a flag
    value indicating whether another Sympson-Hetter iteration is required.  Note
    that number of S-H iterations will not exceed the &maxiter value */

      data comparex;
            set item_par_full;
            keep prob_a1 - prob_a&bins;  /*  Observed exposure rate */
      run;

      data comparex;
            set comparex;
            array over(&bins) over1 - over&bins; /*  Counts number of items exceeding targeted exposure
rate (+ tolerance) */
            array prob_a(&bins) prob_a1 - prob_a&bins;

            retain over1 - over&bins 0;

            /*  Identify items whose probability of administration exceeds the
                  targeted exposure rate (+ tolerance) */
            do i=1 to &bins;
                  if prob_a(i) > &target*(1+&tolerance) then do;
                        over(i) = over(i) + 1;
                        end;
            end;
      run;

      /*  Obtain final tally of items exceeding targeted exposure rate, set
            flag to 1 if another iteration is required */
      data comparex;
            set comparex;
            if _n_ = &nitems;
            array over(&bins) over1 - over&bins;
            sum = 0;
            do i = 1 to &bins;
                  sum = sum + over(i);
            end;
            if sum > 0 then do;
                  call symput('flag',1);
                  end;
            else do;
                  call symput('flag',0);
                  end;
      run;

%mend comparex;

%macro expcon;

      /*  Extract Sympson-Hetter exposure control parameters from the
    ITEM_PAR_FULL dataset */

      data expcon;
            set item_par_full;
            keep expcon1 - expcon&bins;
    run;

      /*  Convert expcon dataset to a column vector */
      data expcon2;
      set expcon;
          array y{*} expcon1-expcon&bins;
```

```
            keep p;
            do j=1 to &bins;
            p=y{j};
            output;
            end;
      run;

      /*  Create a row vector with these exposure control parameters as elements */
      /*  This vector will be passed to the CATLOOP dataset */
      proc transpose data=expcon2 out=expcon prefix=expcon;
            var p;
      run;

      /*  Reset the exposure counters for this iteration */
      data item_par_full;
            set item_par_full;
            array exposed_s(&bins) exposed_s1 - exposed_s&bins;
            array exposed_a(&bins) exposed_a1 - exposed_a&bins;
            do i=1 to &bins;
                  exposed_s(i) = 0;
                  exposed_a(i) = 0;
            end;
      run;

%mend expcon;

%macro icat;
******************************************************************;
*   SUBROUTINE ICAT                                                                          ;
*                                                                 ;
*   INTERACTIVE CAT ADMINISTRATION                                               ;
*   Highest-level module for executing
    an _interactive_  CAT administration ;
******************************************************************;

      %igencat2;        /*  Simulate examinee responses to all items; CAT will select */
      %quad;                  /*  Find quadrature points and assign priors */
      %quadgrid;      /*  Fine-scaled quadrature points */
      /*  Prepare item sets for the CAT administration */
      /* Order of infofn and itemproc important here! */
/*    %infofn;
*/
      %itemproc;
/*    %inforow;
      */

      %expcon;

/************************************************************/
/*        BEGIN CAT ADMINISTRATION                          */
/************************************************************/

      %catloop;

      data cat;
            set catloop;
      run;

%mend icat;

%macro expcat;
*****************************************************************;
/*  SUBROUTINE EXPCAT

      Highest-level module for executing a CAT administration
    where the Sympson-Hetter exposure control parameters are identified
    using iterative simulations

      May identify these parameters conditional or unconditional on estimated
      ability

      INPUT DATASETS:   <none>
      INPUT MACRO VARIABLES:  &NSUBJ, &CATLENGTH
      OUTPUT DATASETS:  CAT
*/
```

```
*****************************************************************;

      %symphettype;     /*  Determine if the procedure will be conditional or unconditional */

      %gencat2;         /*  Simulate examinee responses to all items; CAT will select */
      %quad;                  /*  Find quadrature points and assign priors */

      %quadgrid       /*  Fine-scaled quadrature points and priors */

      /*  Prepare item sets for the CAT administration */
      /* Order of infofn and itemproc important here! */

      %infofn;
      %itemproc;
      %inforow;

/************************************************************/
/*        BEGIN CAT ADMINISTRATION                          */
/************************************************************/

   %let counter = 0;
   %let flag = 1;   /*  Flag = 1 <==>  perform a S-H iteration */

   %do %until (&flag=0);

            %let counter = &counter + 1;

            /*  Append Sympson-Hetter exposure control parameters */
            %expcon;

            /*  Run CAT */
            %catloop;

            data cat;
                  set catloop;
            run;

            /*  Adjust Sympson-Hetter exposure control parameters */
            %symphet;

            /*  Subroutine COMPAREX will generate flag indicating whether another
                  iteration is needed */
            %comparex;

            /*  If current number of iterations exceed &maxiter, exit loop */
            %if (&counter = &maxiter) %then %do;
                  %let flag=0;
                  %end;

   %end;

   /*  Output exposure control parameters */
   %expconout;

%mend expcat;

%macro igencat2;

/*  Generate full pattern of item responses before CAT executes.  This
      method saves computational time in SAS */


%seeds;
%ithetas;
%itempar;

%let nitems=&npool;

%cattype; /*  Set program parameters for fixed- or variable-length CAT */
%mergegen;
%gendata;

data gen_data;
      set gen_data;
      rename p1-p%eval(&nitems * &nparms) = par1-par%eval(&nitems * &nparms);
      rename model1-model&nitems = parmodel1-parmodel&nitems;
      rename ncat1-ncat&nitems = parncat1-parncat&nitems;
      rename ix1-ix&nitems = parix1 - parix&nitems;
      rename theta1 = true1;

run;

%mend igencat2;

%macro ithetas;
/*************************/
/*  SUBROUTINE ITHETAS   */
/*                       ***********************************/
/*  Generate a vector of abilities as specified by input
    file <&ifile> and replicated &nreps times              */

      %global nsubj;

      /*  Load data from <&ifile> */

      data thetas;
            infile wrkdir(&ifile);
            input theta1-theta&ndims;
            call symput('nthetas',_n_);
      run;

      /*  Replicate each observation &nreps times */
      data thetas;
            set thetas;
            do i=1 to &nreps;
                  output;
                  end;
      run;

      /*  Evaluate total number of subjects */
      %let nsubj = %eval(&nthetas * &nreps);

%mend ithetas;
```

```
%macro graphs;

%if &graphs=1 %then %do;

symbol i=spline v=plus;

    /*  This macro works with interactive or non-interactive CAT */
    /*  It produces graphs of the provisional ability estimates versus
        item administered */

      /*  Prepare a dataset with provisional ability estimates */
      data sequence;
            set cat;
            keep prov1-prov&catlength examinee;
            examinee=_n_;
      run;

      proc transpose data=sequence out=graphseq prefix=thtest;
            by examinee;
      run;

      /*  Prepare a dataset with forecasted ability estimates */
      data fsequence;
            set cat;
            keep forecast1-forecast&catlength examinee;
            examinee=_n_;
      run;

      proc transpose data=fsequence out=fgraphseq prefix=forecast;
            by examinee;
      run;

      /*  Prepare a dataset with the true abilities of examinees */

      data true;
            set cat;
            keep true1 examinee;
            examinee=_n_;
      run;

      data true2;
            set true;
            do item=1 to &catlength;
                  output;
            end;
      run;

      /*  Now merge these three datasets */

      data pattern;
            merge graphseq fgraphseq true2;
      run;

      %do g=1 %to &nsubj;

            /*  Pick one examinee at a time */
            data trace;
                  set pattern;
                  if examinee=&g;
                  call symput('ability',true1);
            run;

      %let ability=%sysevalf(&ability); /*  Need to process as floating point */

            /*  Plot provisional estimates against item administration number;
                also plot true ability value */

            proc gplot data=trace;
                  title "Examinee # &g with theta = &ability";
                  plot (true1 thtest1 forecast1)*item /overlay;
            run;

      %end;

%end;
```

```
%mend graphs;
```

APPENDIX D

Exact solution for Fisher interval information (FII)

Define the integral for Fisher interval information as $\int_x^y I_i(\theta)d\theta$, where $I_i(\theta)$ is the information function for item *i*, and *x* and *y* are points on the ability continuum such that $x < y$.

To simplify notation, drop the subscript *i* from the information function. Then the information function is given by

$$I(\theta) = \frac{[P'(\theta)]^2}{P(\theta)[1 - P(\theta)]} \qquad \text{(Eq. 62)}$$

Now let *A* = *Da*, where *a* is the discrimination parameter, *B* = *b*, where *b* is the difficulty parameter, and *C* = *c*, where *c* is the pseudo-guessing parameter. Thus, the 3P logistic model is given by

$$P(\theta) = C + \frac{1 - C}{\{1 + \exp[-A(\theta - B)]\}} \qquad \text{(Eq. 63)}$$

Then the exact solution for $\int_x^y I_i(\theta)d\theta$ under the 3P logistic model is

$$\int_x^y I_i(\theta)d\theta = t_1 - t_2 \qquad \text{(Eq. 64)}$$

where

$$t_1 = \frac{A\left[(C-1)e^{AB} + C\left(e^{AB} + e^{Ax}\right)\ln\left(1 + e^{A(x-B)}\right) - C\left(e^{AB} + e^{Ax}\right)\ln\left(C + e^{A(x-B)}\right)\right]}{(C-1)\left(e^{AB} + e^{Ax}\right)} \qquad \text{(Eq. 65)}$$

and

$$t_2 = \frac{A\left[(C-1)e^{AB} + C\left(e^{AB} + e^{Ay}\right)\ln\left(1 + e^{A(y-B)}\right) - C\left(e^{AB} + e^{Ay}\right)\ln\left(C + e^{A(y-B)}\right)\right]}{(C-1)\left(e^{AB} + e^{Ay}\right)} \qquad \text{(Eq. 66)}$$

BIBLIOGRAPHY

BIBLIOGRAPHY

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* (pp. 397-479). 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.

Davey, T. & Parshall, C.G. (1995). *New algorithms for item selection and exposure control with computerized adaptive testing*. Paper presented at the annual meeting of the American Educational Research Association. April, 1995, San Francisco, CA.

Drasgow, F., Levine, M.V., & Williams, E.A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*, 67-86.

Eignor, D.R., Stocking, M.L., Way, W.D., & Steffen, M. (1993, November). *Case studies in computer adaptive test design through simulation* (Research Report 93-56). Princeton, NJ: Educational Testing Service.

Fan, M. & Hsu, Y. (1996, April). *Utility of Fisher information, global information and different starting abilities in mini CAT.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York.

Kingsbury, G.G. & Zara, A.R. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied Psychological Measurement*, *4*, 241-261.

Lord, F.M. (1977). A broad-range test of verbal ability. *Applied Psychological Measurement*, *1*, 95-100.

Lord, F.M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika*, *48*, 233-245.

Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

McBride, J.R. & Martin, J.T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D.J. Weiss (Ed.), *New horizons in testing*. New York: Academic Press.

Nering, M.L. (1997). The distribution of indexes of person fit within the computerized adaptive testing environment. *Applied Psychological Measurement*, *21*, 115-127.

Nering, M.L. & Meijer, R.R. (1998). A comparison of the person response function and the l_z person-fit statistic. *Applied Psychological Measurement*, *22*, 53-69.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*(4, pt. 2).

SAS Institute (2000). The SAS System for Windows, Release 8.01. Cary, NC: SAS Institute, Inc.

Spanos, A. (1999). *Probability theory and statistical inference: Econometric modeling with observational data*. Cambridge, England: Cambridge University Press.

Stocking, M.L. (1993). *Controlling item exposure rates in a realistic adaptive testing paradigm* (Research Report 93-2). Princeton, NJ: Educational Testing Service.

Stocking, M.L. & Lewis, C. (1995, August). *A new method of controlling item exposure in computerized adaptive testing* (Research Report 95-25). Princeton, NJ: Educational Testing Service.

Stocking, M.L. & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, *23*, 57-75.

Stocking, M.L. & Lewis, C. (2000). Methods of controlling the exposure of items in CAT. In In W.J. van der Linden& C.A.W. Glas (Eds.), *Computerized adaptive testing: theory and practice*. Dordrecht, The Netherlands: 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.

Stocking, M.L. & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement*, *17*, 277-292.

Stocking, M.L. & Swanson, L. (1996, December). *Optimal design of item pools for computerized adaptive tests* (Research Report 96-34). Princeton, NJ: Educational Testing Service.

Suen, H.K. (1990). *Principles of test theories*. Mahwah, NJ: Lawrence Erlbaum Associates.

Swanson, L. & Stocking, M.L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement*, *17*, 151-166.

Sympson, J.B. & Hetter, R.D. (1985). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military Testing Association*. San Diego, CA: Navy Personnel Research and Development Center.

Tatsuoka, K.K. (1984). Caution indices based on item response theory. *Psychometrika*, *49*, 95-110.

Thissen, D. & Mislevy, R.J. (1990). Testing algorithms. In H. Wainer, *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Trabin, T.E. & Weiss, D.J. (1983). The person response curve: Fit of individuals to item characteristic curve models. In D.J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 83-108). New York: Academic Press.

Urry, V. W. (1970). *A monte carlo investigation of logistic test models*. Unpublished doctoral dissertation, Purdue University, West Lafayette, IN.

van der Linden,W.J. & Pashley, P.J. (2000). Item selection and ability estimation in adaptive testing. In W.J. van der Linden& C.A.W. Glas (Eds.), *Computerized adaptive testing: theory and practice*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

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.

Wainer, H. & Mislevy, R.J. (1990). Item response theory, item calibration and proficiency estimation. In H. Wainer, *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wang, T. & Vispoel, W.P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*, 109-135.

Warm, T.A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54, 427-450.

Xiao, B. (1999). Strategies for computerized adaptive grading testing. *Applied psychological measurement*, *23*, 136.146.