A Mixture Rasch Model–Based Computerized Adaptive Test for Latent Class Identification
Hong Jiao, George Macready, Junhui Liu and Youngmi Cho

Applied Psychological Measurement 2012 36: 469 originally published online 29 June 2012
DOI: 10.1177/0146621612450068

The online version of this article can be found at:
http://apm.sagepub.com/content/36/6/469

Published by:
SAGE
http://www.sagepublications.com

Additional services and information for Applied Psychological Measurement can be found at:
Email Alerts: http://apm.sagepub.com/cgi/alerts
Subscriptions: http://apm.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav
Citations: http://apm.sagepub.com/content/36/6/469.refs.html

>> Version of Record - Jul 31, 2012
OnlineFirst Version of Record - Jun 29, 2012

What is This?
A Mixture Rasch Model–Based Computerized Adaptive Test for Latent Class Identification

Hong Jiao¹, George Macready¹, Junhui Liu¹, and Youngmi Cho¹

Abstract
This study explored a computerized adaptive test delivery algorithm for latent class identification based on the mixture Rasch model. Four item selection methods based on the Kullback–Leibler (KL) information were proposed and compared with the reversed and the adaptive KL information under simulated testing conditions. When item separation was large, all item selection methods did not differ evidently in terms of accuracy in classifying examinees into different latent classes and estimating latent ability. However, when item separation was small, two methods with class-specific ability estimates performed better than the other two methods based on a single latent ability estimate across all latent classes. The three types of KL information distributions were compared. The KL and the reversed KL information could be the same or different depending on the ability level and the item difficulty difference between latent classes. Although the KL information and the reversed KL information were different at some ability levels and item difficulty difference levels, the use of the KL, the reversed KL, or the adaptive KL information did not affect the results substantially due to the symmetric distribution of item difficulty differences between latent classes in the simulated item pools. Item pool usage and classification convergence points were examined as well.

Keywords
the mixture Rasch model, computerized adaptive test, Kullback–Leibler information, latent class identification

Computerized adaptive testing (CAT) is a class of test delivery algorithms for scoring examinees to achieve higher measurement accuracy and test delivery efficiency. Under this approach, each examinee is provided with an individualized set of test items that have been successively selected and adapted to the current estimate of his or her ability level. The use of CAT in operational testing programs has been advanced due to the innovations in measurement theory and computer technology. The recent adoption of CAT in large-scale state testing programs is a good example of the increasing use of this assessment approach in practice.

¹University of Maryland, College Park, USA

Corresponding Author:
Hong Jiao, Measurement, Statistics and Evaluation, Department of Human Development and Quantitative Methodology, University of Maryland, 1230B Benjamin Building, College Park, MD 20742, USA
Email: hjiao@umd.edu
CAT may be used to obtain point estimates of examinees’ latent ability or to classify examinees into ordered categories along a latent ability scale. Multiple CAT algorithms for classification purposes have been proposed, such as CAT with point estimates for classification (Kingsbury & Weiss, 1983), computerized mastery test using the Bayesian decision theory (Lewis & Sheehan, 1990; Sheehan & Lewis, 1992), and sequential probability ratio test (Wald, 1947)–based computerized classification testing (Eggen, 1999; Ferguson, 1969; Reckase, 1983; Spray, 1993; Thompson, 2009). These CAT classification algorithms rely on cut scores from a standard setting procedure for classification decisions.

Classification decisions could also be made using a latent category model–based approach. Such approaches are connected in one way or another to latent class analysis (Templin & Jiao, 2011). It could be solely based on a latent class modeling framework (e.g., Brown, 2007; Macready & Dayton, 1992; Templin, Poggio, Irwin, & Henson, 2007) or a mixture item response theory (IRT) model. This latter modeling approach is a derivative of latent class modeling in which a latent class variable is crossed with a continuous latent ability variable like in the Rasch model (e.g., Jiao, Lissitz, Macready, Wang, & Liang, 2011; Templin, Cohen, & Henson, 2008). This study explores a CAT algorithm to facilitate classification of examinees into known latent classes based on responses to items adaptively selected from an item pool that has been calibrated using the mixture Rasch model.

The Mixture Rasch Model

The mixture Rasch model (Kelderman & Macready, 1990; Mislevy & Verhelst, 1990; Rost, 1990; von Davier & Rost, 2006) integrates the Rasch model and the latent class model to detect the possible presence of multiple latent populations underlying item responses. In the Rasch model, a continuous latent variable underlies an item performance whereas in the latent class model, the latent class membership underlies an item performance. The mixture Rasch model assumes that examinees are from multiple latent populations and that the Rasch model holds within each population (i.e., latent class). However, the item difficulty parameters differ across latent classes. The combination of the Rasch model and the latent class model allows for simultaneous estimation of a continuous latent ability that provides a measure of the trait of interest, as well as a discrete latent group membership which also differentiates examinees in their likelihood of a correct item response. The continuous quantitative variable can be interpreted in the same way as the latent ability parameter in the Rasch model. The discrete qualitative variable can be related to such factors as problem-solving strategies (Mislevy & Verhelst, 1990) or test speededness (Bolt, Cohen, & Wollack, 2002).

In the mixture Rasch model, the conditional probability of a correct response for examinee $j$ to item $i$ is expressed in Equation 1:

$$P_{jic} = \frac{1}{1 + \exp \left[-(\theta_{jc} - b_{ic})\right]},$$

where $P_{jic}$ is the probability of the $j$th examinee with latent ability $\theta_{jc}$ in latent class $c$ responding correctly, $x_i = 1$, to the $i$th item with difficulty $b_{ic}$ for that particular latent class $c$. It is assumed that $\theta_{jc}$ is a continuous latent variable. Each examinee’s latent class membership is $c$ and labeled in the ability parameter as $\theta_{jc}$.

The unconditional probability of a correct response to item $i$ across latent classes is expressed in Equation 2:
$P_{ji} = \sum_c \pi_c p_{jic}$

where $P_{ji}$ is the unconditional probability of a correct item response and $\pi_c$ is the class mixing proportion (or the percentage of examinees in a specific latent class $c$), with constraints $0 < \pi_c < 1$ and $\sum_c \pi_c = 1$ across classes.

Each examinee’s latent class membership is assessed by comparing the posterior probabilities of that examinee being a member of each latent class. An examinee is assigned to the latent class with the highest posterior probability. The posterior probability for each latent class given an examinee’s latent ability, item difficulty parameters for each latent class, and a response vector is presented in Equation 3 (the notations in Equation 3 follow the same interpretations as those in Equations 1 and 2):

$P_{jc} = \frac{\sum_c \pi_c \prod_i (P_{jic})^x (1 - P_{jic})^{1-x}}{\prod_c \sum \pi_c \prod_i (P_{jic})^x (1 - P_{jic})^{1-x}}.

Over the last two decades, there has been an increase in the applications of the mixture IRT models in addressing a variety of psychometric issues. For instance, the mixture Rasch model was proposed in identifying latent classes utilizing different cognitive strategies in solving problems based on different item response patterns (Mislevy & Verhelst, 1990; Rijmen & De Boeck, 2003; Rost & von Davier, 1993). Differential item functioning across latent classes was investigated by comparing different item parameter estimates across latent classes (Cohen & Bolt, 2005; De Ayala, Kim, Stapleton, & Dayton, 2002; Kelderman & Macready, 1990). Different latent subpopulations caused by test speededness can be identified by data analysis using the mixture IRT models (Bolt et al., 2002; Boughton & Yamamoto, 2004; Yamamoto, 1989; Yamamoto & Everson, 1997). All these studies utilize the information embedded in item response patterns to estimate the latent grouping of examinees. These applications count on accurate identification of latent class membership. This study explores whether a CAT based on the mixture Rasch model would provide high classification accuracy of examinees into different latent classes.

There are multiple scenarios where mixture IRT models may find utility. The first scenario is to identify latent class membership of examinees while accounting for the noise variance due to a continuous latent ability which is not explained in traditional latent class models. Under this scenario, the purpose is to increase the classification accuracy. The second scenario is to estimate a continuous latent ability while accounting for local dependence induced by the discrete latent classes which is not explained by the traditional IRT models. Under this scenario, the purpose is to increase the ability parameter estimation accuracy. The third scenario is a combination of both purposes. Under this last scenario, classification accuracy and ability parameter estimation accuracy will be of interest. Each of the scenarios warrants its own line of research. The present study focuses on the first scenario, classification of examinees into latent classes.

The Implementation of a Mixture Rasch Model-Based CAT

The implementation of a CAT requires multiple steps. Wainer (1990) stated the key questions in administering a CAT as follows:

1. How to select an item to start the testing process?
2. How to select subsequent items to be administered after observing an examinee’s responses to the current and the previously administered items?

3. How to establish a stopping rule that specifies when the testing process should be terminated for each examinee?

A CAT may start in different ways (Parshall, Spray, Kalohn, & Davey, 2002). The “best guess” method administers the first item of medium difficulty. The “use what you’ve got” method makes use of other test scores or auxiliary information to obtain an estimate of an examinee’s ability and then selects an item with difficulty most appropriate based on the estimated ability. The “start easy” method begins the testing process with relatively easy items to give examinees time to warm up. In the present study, the mixture Rasch model–based CAT starts by selecting five items randomly from the calibrated item pool. This approach is used because under the mixture Rasch model, some items may be easy for examinees from one latent class but difficult for those who are members of another class. With no information or estimate available about the examinee’s latent class membership, it may not always be possible to pick an easy or medium difficult item for examinees.

After the collection of an examinee’s responses to the initial items, an estimate of the latent class membership is made if needed. The examinee’s latent ability can be estimated using either a maximum likelihood estimation (MLE) method or a Bayesian estimation method such as Bayes’s mean (or expected a posteriori [EAP]) or Bayes’s mode (Baker, 1992) conditional on the latent class membership. In the present study, the EAP method is used for ability estimation because of the relatively short test length explored, which results in relatively high occurrence probabilities of zero or perfect scores, and there is no MLE estimate for such scores. If needed, based on the current ability estimate, the latent class membership is updated by assigning the examinee to the latent class with the highest posterior probability which is computed based on Equation 3.

Usually the next item to be administered to a given examinee will be selected to satisfy three requirements: (a) maximizing test efficiency, (b) content balancing, and (c) item exposure rate control. In the conventional CAT, items can be selected based on maximum Fisher information. In this study, an item may be selected to maximize the latent class separation or to simultaneously maximize the distinction between the latent classes and the ability estimate(s) based on one of the adapted Kullback–Leibler (KL) information indices proposed in the following.

After the administration of the selected item, the examinee’s latent ability is estimated and the examinee’s latent class membership is updated (if needed) based on the posterior probability for each latent class. This process is repeated iteratively until the “stop criterion” is met. In this study, the fixed length method is used to terminate the CAT. As an initial investigation, no content balance or item exposure control is considered.

**The KL Information**

In this study, items are selected based on the maximum KL (Cover & Thomas, 1991) information of the unused items in the item bank. The KL information was originally proposed to measure the distance between two probability distributions specified by the true and the estimated parameters (Cover & Thomas, 1991). Chang and Ying (1996) first proposed the selection of items based on the KL information in conjunction with a unidimensional IRT model–based CAT. The KL information as a global information index was recommended for item selection over the local Fisher information to increase efficiency and accuracy at the initial stage of CAT when the ability estimates may be far away from the true parameter values (Chang & Ying, 1996). The KL information was later applied for item selection in computerized classification
testing using the sequential probability ratio test (Eggen, 1999). More recently, the KL information has been applied in other test settings, such as multidimensional CAT (e.g., Wang, Chang, & Boughton, 2010), CAT for cognitive diagnosis (e.g., McGlohen & Chang, 2008; Xu, Chang, & Douglas, 2003), and test construction for classification in standard setting (e.g., Templin et al., 2008). When classification is the focus of the test, it seems reasonable to use the KL information to distinguish between the two adjacent latent classes.

The KL information function is defined as (Chang & Ying, 1996)

\[ KL_i(\theta_1 \| \theta_2) = E_{\theta_2} \log \frac{L(\theta_1; x_i)}{L(\theta_2; x_i)}, \]  

(4)

where \( KL_i(\theta_1 \| \theta_2) \) is the item KL information for item \( i \) to distinguish between the likelihood functions at two theta points, \( \theta_1 \) and \( \theta_2 \). \( E_{\theta_2} \) is the expected value of the log ratio of the two likelihood functions relative to \( \theta_2 \), which is considered to be the true value. The KL information is a measure of the difference between the two likelihood functions at \( \theta_2 \) and \( \theta_1 \). When the KL information is maximized, the difference between the likelihood functions under the two hypotheses is also maximized. The larger the KL information, the more efficient the test will be. For dichotomous IRT models, the KL item information can be calculated as (Chang & Ying, 1996; Eggen, 1999)

\[ KL_i(\theta_1 \| \theta_2) = P_i(\theta_2) \log \frac{P_i(\theta_2)}{P_i(\theta_1)} + Q_i(\theta_2) \log \frac{Q_i(\theta_2)}{Q_i(\theta_1)}, \]  

(5)

where \( P_i(\theta_1) \) and \( P_i(\theta_2) \) are the probabilities of a correct response to item \( i \) at latent ability levels of \( \theta_1 \) and \( \theta_2 \), respectively, based on an IRT model. \( Q_i(\theta_1) = 1 - P_i(\theta_1) \) and \( Q_i(\theta_2) = 1 - P_i(\theta_2) \) are the probabilities of an incorrect response to item \( i \) at \( \theta_1 \) and \( \theta_2 \), respectively. The KL information is adapted in this study as described in the following section.

The Adapted KL Information

To fully understand the adapted KL information proposed in this study, an elaboration is needed on the nature and the number of the ability parameter estimated in the mixture IRT model.

Latent classes in the mixture IRT modeling could measure the same or different underlying trait(s). As stated in Cohen and Bolt (2005) and Rost (1990), the same latent trait could be measured but with different distributional properties specified by the central location and the variability across latent classes. This is essentially a multigroup unidimensional IRT model with the group membership to be estimated. However, as stated in Mislevy and Verhelst (1990), the latent traits measured across latent classes may be of different natures. This is essentially a multigroup multidimensional IRT model. An examinee’s latent group membership is to be estimated. The latent trait measured within each latent class is unidimensional but the latent traits measured across latent classes are multidimensional. However, an examinee’s response to an item is still governed by the within-group unidimensional latent trait conditional on the latent class membership. In this study, all item selection methods are proposed under the assumption that the same latent trait is measured across latent classes.

Assuming an item pool has been calibrated using the mixture Rasch model with two latent classes and the item parameter estimates are obtained from the prior calibration, the current literature documented two methods in the estimation of ability parameters in the mixture IRT modeling. One method estimates one ability parameter for each examinee across latent classes, whereas the other method estimates two class-specific ability parameters for each examinee. An example of the former procedure is the Bayesian Markov chain Monte Carlo estimation of the
mixture IRT models, whereas the latter method can be found in the Multidimensional Discrete Latent Trait Model (mdltm) software (von Davier, 2005) using the marginal MLE method with the expectation–maximization algorithm. Both methods have been applied in estimating ability parameters in the mixture Rasch model.

When one single latent ability parameter is estimated across latent classes, the CAT algorithm works as follows in the present study. After the administration and scoring of five randomly selected items for an examinee, an initial assignment of the examinee to a latent class is made based on the number of items he or she answers correctly. If the examinee answers three or more items correctly, he or she will be assigned to the latent class with a higher mean ability level (assume latent classes with ordered ability distributions in this study although it is possible to assume equal mean ability for different classes). Otherwise, the initial assignment of the latent class membership will be the class with a lower mean latent ability. After the initial assignment of the latent class membership, the latent ability is estimated. Given the ability estimate, the latent class membership is updated based on the posterior probability for each latent class. An examinee’s latent class membership and latent ability are separately estimated, conditional on the estimate of the other quantity. The rationale behind this procedure is that every examinee only has one ability parameter, but the likelihood that an examinee is in each latent class could be estimated using the item response pattern observed for that examinee.

When class-specific ability parameters are estimated for each examinee, the CAT algorithm works differently. A class-specific ability parameter is estimated for each latent class after administering items adaptively selected based on one of the proposed KL information. There is no interim estimation of the latent class membership. When the maximum test length is reached, the posterior probability for each latent class is obtained for each ability estimate. Then the latent class membership is determined by comparing the sum of the posterior probabilities for each class based on each class-specific ability estimate. The rationale for this method is that although in theory every examinee has a single ability parameter, the probability that a given examinee is classified into each latent class is rarely zero. Without knowing which class-specific ability estimate is closer to the true parameter, the use of an ability estimate from each latent class may provide more information regarding the likelihood of the latent class membership.

Both methods are investigated in this study. Related to each method of ability estimation and latent class updating, the KL information can be adapted differently as follows. To simplify the explanation and demonstration, this study assumes two latent classes where Class 2 has a higher mean ability than does Class 1, but the proposed methods can be extended to conditions where there are more than two latent classes as well as when latent classes are not ordered.

**Method 1**

Assuming estimation of a single latent ability across all latent classes, to maximize the KL information between two latent classes at the current ability estimate, the KL information can be adapted as

\[
KL_i(c_1||c_2) = P_i(c_2, \hat{\theta}) \log \frac{P_i(c_2, \hat{\theta})}{P_i(c_1, \hat{\theta})} + Q_i(c_2, \hat{\theta}) \log \frac{Q_i(c_2, \hat{\theta})}{Q_i(c_1, \hat{\theta})},
\]

where \(P_i(c_1, \hat{\theta})\) and \(P_i(c_2, \hat{\theta})\) are the conditional probabilities of a correct response to item \(i\) for Latent Classes 1 and 2, respectively, at the current ability estimate \(\hat{\theta}\). \(Q_i(c_1, \hat{\theta}) = 1 - P_i(c_1, \hat{\theta})\) and \(Q_i(c_2, \hat{\theta}) = 1 - P_i(c_2, \hat{\theta})\) are the conditional probabilities of an incorrect response. This algorithm maximizes the information to distinguish between the latent classes conditional on the
current ability estimate. This method is appropriate for use when the same latent ability is measured across latent classes.

**Method 2**

Also assuming estimation of a single latent ability across all latent classes, the KL information can be adapted to maximize the distinction between latent classes as well as between the current ability estimate and its true value. According to Chang and Ying (1996), when the test is used for the point estimation of the ability parameter, a single KL information index can be obtained by taking the average over an appropriate interval for the current ability estimate. The lower and the upper bounds of the interval can be defined as \( \hat{\theta}_L = \hat{\theta} - \delta \) and \( \hat{\theta}_U = \hat{\theta} + \delta \), where \( \delta = c/\sqrt{n} \), \( n \) is the number of items previously administered to a specific examinee on which his or her current ability estimate is based, and \( c \) is a constant selected according to a specified coverage probability. For example, if a 95% confidence interval around the current ability estimate is specified, \( c \) can be set to be 1.96. This is the value used in the present study. Although Chang and Ying suggested taking the average over an appropriate interval of the current ability estimate, the use of endpoints is adopted here in obtaining a single index of the KL information following Eggen 1999, which used the KL information for classification. The adapted KL information can be expressed as in Equation 7:

\[
\text{KL}_i(c_1, \hat{\theta}_L \mid c_2, \hat{\theta}_U) = P_i(c_2, \hat{\theta}_U) \log \frac{P_i(c_2, \hat{\theta}_U)}{P_i(c_1, \hat{\theta}_L)} + Q_i(c_2, \hat{\theta}_U) \log \frac{Q_i(c_2, \hat{\theta}_U)}{Q_i(c_1, \hat{\theta}_L)},
\]

where \( P_i(c_1, \hat{\theta}_L) \) and \( P_i(c_2, \hat{\theta}_U) \) are the joint probabilities of a correct response to item \( i \) for Latent Class 1 at the lower bound and Latent Class 2 at the upper bound of the interval around the current ability estimate, assuming that Class 2 has a higher mean ability than does Class 1. \( Q_i(c_1, \hat{\theta}_L) = 1 - P_i(c_1, \hat{\theta}_L) \) and \( Q_i(c_2, \hat{\theta}_U) = 1 - P_i(c_2, \hat{\theta}_U) \) are the joint probabilities of an incorrect response. This algorithm maximizes the information to distinguish between both latent classes and the upper and lower bounds of the interval set around the current ability estimate. Like Method 1, this method is appropriate for use when the same latent ability is measured across latent classes.

**Method 3**

Assuming estimation of one latent ability for each latent class (each examinee has two class-specific ability estimates if the number of classes is two), the KL information can further be adapted to maximize the distinction between latent classes and between current ability estimates for each latent class. The adapted KL information can be expressed as in Equation 8:

\[
\text{KL}_i(c_1, \hat{\theta}_1 \mid c_2, \hat{\theta}_2) = P_i(c_2, \hat{\theta}_2) \log \frac{P_i(c_2, \hat{\theta}_2)}{P_i(c_1, \hat{\theta}_1)} + Q_i(c_2, \hat{\theta}_2) \log \frac{Q_i(c_2, \hat{\theta}_2)}{Q_i(c_1, \hat{\theta}_1)},
\]

where \( P_i(c_1, \hat{\theta}_1) \) and \( P_i(c_2, \hat{\theta}_2) \) are the joint probabilities of a correct response to item \( i \) for Latent Classes 1 and 2 at the current class-specific ability estimates for their respective latent classes. \( Q_i(c_1, \hat{\theta}_1) = 1 - P_i(c_1, \hat{\theta}_1) \) and \( Q_i(c_2, \hat{\theta}_2) = 1 - P_i(c_2, \hat{\theta}_2) \) are the joint probabilities of an incorrect response.

There is no interim latent class membership updating. An examinee’s latent class membership is estimated only after the administration of the last item by comparing the posterior probabilities for his or her membership in each latent class. Due to the estimation of class-specific latent ability,
the posterior probability of membership in each latent class is related to each class-specific ability estimate. Thus, the posterior probability for membership in Latent Class 1 is the sum of the posterior probability of membership in Latent Class 1 based on Latent Class 1 ability estimate and the posterior probability of membership in Latent Class 1 based on Latent Class 2 ability estimate, whereas the posterior probability for membership in Latent Class 2 is the sum of the posterior probability of membership in Latent Class 2 based on Latent Class 1 ability estimate and the posterior probability of membership in Latent Class 2 based on Latent Class 2 ability estimate.

Method 3 may be used when either the same or different latent ability is measured across latent classes. However, when different latent traits are assumed to be measured across latent classes, the posterior probability for membership in each latent class should be computed based on the class-specific ability in its own class only. That is, the posterior probability for membership in Latent Class 1 is the posterior probability of membership in Latent Class 1 based on Latent Class 1 ability estimate, whereas the posterior probability for membership in Latent Class 2 is the posterior probability of membership in Latent Class 2 based on Latent Class 2 ability estimate.

Method 4
Method 4 is essentially an extension of Methods 1 and 3. Although a single latent ability is estimated in Method 1, the probability that a person is in the other class is never zero. Items are selected to maximize the ratio of the two class-specific probabilities conditional on the single ability estimate. However, although Method 4 estimates class-specific abilities like Method 3, only one will be the ability parameter underlying item performance. Without knowing which class-specific ability estimate is closer to the true value, it is essentially assumed that each will be a possibility.

Following the logic in Method 1, Method 4 treats each class-specific ability estimate as the sole current ability estimate. The KL information can be computed based on each class-specific ability estimate. The posterior probability for each latent class can be used as the weight in the KL information summation. If the posterior probability for Latent Class 1 is higher than that for Latent Class 2, the KL information based on the ability estimate for Latent Class 1 takes on more weight, and vice versa. The KL information is a sum of the weighted KL information assuming each class-specific ability estimate is the current ability estimate. The weighted KL information based on each class-specific ability estimate makes use of all possible sources of information. This method is only appropriate for use when the same latent trait is measured across the two classes.

Method 4 works as follows. Assuming estimation of class-specific latent ability for each latent class, this method involves weighting the KL information to maximize the distinction between latent classes. The KL information computed based on Class 1 ability estimate is given in Equation 9 and that based on Class 2 ability estimate is given in Equation 10:

\[
KL_{i \hat{\theta}_1}(c_1||c_2) = P_i(c_2, \hat{\theta}_1) \log \frac{P_i(c_2, \hat{\theta}_1)}{P_i(c_1, \hat{\theta}_1)} + Q_i(c_2, \hat{\theta}_1) \log \frac{Q_i(c_2, \hat{\theta}_1)}{Q_i(c_1, \hat{\theta}_1)},
\]

(9)

\[
KL_{i \hat{\theta}_2}(c_2||c_1) = P_i(c_2, \hat{\theta}_2) \log \frac{P_i(c_2, \hat{\theta}_2)}{P_i(c_1, \hat{\theta}_2)} + Q_i(c_2, \hat{\theta}_2) \log \frac{Q_i(c_2, \hat{\theta}_2)}{Q_i(c_1, \hat{\theta}_2)}.
\]

(10)

The new terms in the above equations are noted as follows: \(P_i(c_2, \hat{\theta}_1)\) is the joint probability of a correct response to item \(i\) for Latent Class 2 at the current ability estimate for Latent Class 1, and \(P_i(c_1, \hat{\theta}_2)\) is the joint probability of a correct response to item \(i\) for Latent Class 1 at the current ability estimate for Latent Class 2. Correspondingly, \(Q_i(c_2, \hat{\theta}_1) = 1 - P_i(c_2, \hat{\theta}_1)\)
and \( Q_1(c_1, \hat{\theta}_2) = 1 - P_1(c_1, \hat{\theta}_2) \) are the joint probabilities of an incorrect response. Item selection is based on the sum of the weighted KL information expressed in Equations 9 and 10. For the KL information computed in Equation 9, the weight is the average of the summed posterior probability for Latent Class 1 based on each class-specific ability estimate, whereas the weight for the KL information computed in Equation 10 is the average of the summed posterior probability for Latent Class 2 based on each class-specific ability estimate. The same method as described in Method 3 is used to compute the posterior probabilities. The latent class membership is estimated only after the administration of the last item.

As indicated in Chang and Ying (1996), the KL convergence is not a symmetric quantity; that is, \( KL_1(c_1||c_2) \neq KL_2(c_2||c_1) \). In the indices proposed above, the KL information is relative to \( c_1||c_2 \). This study also explores the situation where Latent Classes 1 and 2 are switched as \( KL_2(c_2||c_1) \) in Equations 6 through 10 and labels them as the reversed KL information. Furthermore, as suggested by one reviewer, a new index is defined in terms of \( \sum KL(c_i||\hat{c}) \), where \( \hat{c} \) is the current estimate of an examinee’s latent class. When \( c_i = \hat{c} \), \( KL(c_i||\hat{c}) = 0 \). Such an index would indicate the degree of separation to which a given item reveals when the current latent class estimate is incorrect; indeed the examinee is a member of another latent class. This set of indices is explored as well and labeled as the adaptive KL information in the present study.

**Illustrations With Simulation**

The four proposed item selection methods based on the KL information in the mixture Rasch model–based CAT were illustrated in simulation studies where two latent classes were simulated. Two latent classes were frequently studied in the mixture IRT models (Li, Cohen, Kim, & Cho, 2009; Rost, 1990). Furthermore, the reversed KL and the adaptive KL indices were compared with the proposed four KL indices. All together, 12 item selection methods were compared within each of the simulated four item pools.

**Item Pool Construction and Calibration**

Four item pools were simulated each with 500 items. The characteristics of item pools were delineated by the item difficulty and ability distributions in each latent class. For Scenario 1 labeled as large item separation, item difficulty parameters were simulated from a standard normal distribution for Latent Class 1, whereas those for Latent Class 2 took on the negative values of those generated for Latent Class 1 (i.e., the correlation between item difficulties for the two latent classes was \(-1\)). This resembles the latent class structure simulated in the Rost (1990) study and represents a best-case scenario with distinct separation of latent classes in terms of item difficulty. For Scenario 2 labeled as small item separation, item difficulty parameters were the same as those simulated in Scenario 1 for Latent Class 1, whereas those for Latent Class 2 were more difficult for the first half of the items by 1 logit unit and easier for the other half by \(-1\) unit. This resembles the latent class structure simulated in the Li et al. (2009) study.

Ability parameters for Latent Class 1 were simulated from a standard normal distribution, whereas those for Latent Class 2 were simulated from a normal distribution with a mean of 1 and a standard deviation of 1 in Condition 1 labeled as large ability separation. In Condition 2 labeled as small ability separation, ability parameters for Latent Class 1 were simulated from a standard normal distribution, whereas those for Latent Class 2 were simulated from a normal distribution with a mean of 0.5 and a standard deviation of 1. By fully crossing the levels of item difficulty and ability parameter distributions, four item pools were constructed. In Pool 1, large item difficulty separation was paired with large ability separation. In Pool 2, small item...
separation was paired with large ability separation. In Pool 3, large item separation was paired with small ability separation. In Pool 4, small item separation was paired with small ability separation.

Item responses were generated based on Equation 1 using the true item difficulty and ability parameters for 10,000 examinees. The mixing proportion was 50% for both latent classes. Items were calibrated under a testing scenario where each examinee answered 20 items randomly selected from the item pool. This mimics 96% missing responses often encountered in real-world CAT item pool calibration. The mdltm software (von Davier, 2005) was used for calibrating items. For model identification, the sum of item difficulties was set at 0 for both latent classes.

Simulation of the Mixture Rasch Model–Based CAT

To simulate a CAT, ability parameters were generated for 10,000 examinees with 5,000 examinees in each latent class from the same ability distributions as used for the respective item pool calibration. In the CAT simulation, item responses were generated based on an examinee’s true ability and true item difficulty parameters from the true latent class that the examinee belongs to. For Methods 1 and 2 with a single ability estimate across classes, after the administration of the first 5 items randomly selected from the item pool, each examinee’s ability parameter was estimated based on the temporary assignment to a latent class for that examinee following the majority rule (the number of correct responses was larger than three). Then, the latent class membership was updated and the examinee’s ability parameter was estimated alternatingly after the administration of each item adaptively selected based on one of the KL information indices. For Methods 3 and 4 with class-specific ability estimates, only class-specific ability parameters were estimated after the administration of the first 5 items randomly selected and updated after administering each subsequent item adaptively selected. When the last item was administered, testing was terminated and the final estimates were updated for the ability parameter and the latent class membership for each examinee. The 12 item selection methods were compared in terms of classification accuracy, the absolute bias in the ability parameter estimation, and the correlation between the estimated and the true ability parameters.

Results

The results for Pool 1 are summarized in Table 1. The classification accuracy among the four methods did not differ much for this pool. Methods 1 and 2 produced slightly lower classification accuracy, lower correlations between the true and the estimated ability parameters, and slightly higher mean absolute bias in the ability parameter estimates than did Methods 3 and 4. The KL, the reversed KL, and the adaptive KL information had little impact on classification accuracy under each of the four methods. The correlations between the true and the estimated ability parameters were generally close irrespective of the use of the KL or the reversed KL information but higher for the adaptive KL information. The mean absolute bias in the ability parameter estimates was about the same for the KL and the reversed KL information, but lower for the adaptive KL information.

To better understand the proposed methods for item selection, 20 items were randomly selected to construct a nonadaptive linear form for each examinee. There was no adaptation in item selection in this random linear form construction. As presented in Table 1, classification accuracy was reduced, but the ability parameter estimation accuracy improved with smaller mean absolute bias and higher correlations between the true and the estimated parameters.
In general, the discrepancies in the evaluation criteria observed under item Pool 1 were not large across the proposed item selection methods. This outcome may be a result of the characteristics of this item pool. The mean ability difference for the two latent classes was about 1 logit unit. The largest absolute difference in the item difficulty estimates was about 5.9 logit units. Although only a small percentage of items had such large differences, those items were the most frequently used. The separation delineated by the ability and item difficulty distributions between the two latent classes was relatively large.

The results for Pool 2 are presented in Table 2. Compared with Pool 1, the classification accuracy decreased dramatically. This may be due to the different distributions of item difficulty

### Table 1. Estimation of Ability Parameters and Classification Accuracy for Pool 1

<table>
<thead>
<tr>
<th>KL</th>
<th>Item selection algorithm</th>
<th>Absolute bias</th>
<th>Correlation</th>
<th>Classification accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>KL, KL(c1</td>
<td></td>
<td>c2)</td>
<td>Method 1</td>
<td>0.462</td>
</tr>
<tr>
<td></td>
<td>Method 2</td>
<td>0.457</td>
<td>.372</td>
<td>.853</td>
</tr>
<tr>
<td></td>
<td>Method 3</td>
<td>0.446</td>
<td>.337</td>
<td>.869</td>
</tr>
<tr>
<td></td>
<td>Method 4</td>
<td>0.455</td>
<td>.337</td>
<td>.850</td>
</tr>
<tr>
<td>Reversed KL, KL(c2</td>
<td></td>
<td>c1)</td>
<td>Method 1</td>
<td>0.482</td>
</tr>
<tr>
<td></td>
<td>Method 2</td>
<td>0.506</td>
<td>.450</td>
<td>.802</td>
</tr>
<tr>
<td></td>
<td>Method 3</td>
<td>0.447</td>
<td>.336</td>
<td>.867</td>
</tr>
<tr>
<td></td>
<td>Method 4</td>
<td>0.463</td>
<td>.358</td>
<td>.853</td>
</tr>
<tr>
<td>Adaptive KL, ∑KL(c1</td>
<td></td>
<td>c)</td>
<td>Method 1</td>
<td>0.447</td>
</tr>
<tr>
<td></td>
<td>Method 2</td>
<td>0.478</td>
<td>.426</td>
<td>.832</td>
</tr>
<tr>
<td></td>
<td>Method 3</td>
<td>0.440</td>
<td>.331</td>
<td>.873</td>
</tr>
<tr>
<td></td>
<td>Method 4</td>
<td>0.418</td>
<td>.331</td>
<td>.884</td>
</tr>
<tr>
<td></td>
<td>Random</td>
<td>0.388</td>
<td>.303</td>
<td>.898</td>
</tr>
</tbody>
</table>

Note: KL = Kullback–Leibler.

### Table 2. Estimation of Ability Parameters and Classification Accuracy for Pool 2

<table>
<thead>
<tr>
<th>KL</th>
<th>Item selection algorithm</th>
<th>Absolute bias</th>
<th>Correlation</th>
<th>Classification accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>KL, KL(c1</td>
<td></td>
<td>c2)</td>
<td>Method 1</td>
<td>0.395</td>
</tr>
<tr>
<td></td>
<td>Method 2</td>
<td>0.463</td>
<td>.358</td>
<td>.856</td>
</tr>
<tr>
<td></td>
<td>Method 3</td>
<td>0.366</td>
<td>.275</td>
<td>.914</td>
</tr>
<tr>
<td></td>
<td>Method 4</td>
<td>0.377</td>
<td>.278</td>
<td>.908</td>
</tr>
<tr>
<td>Reversed KL, KL(c2</td>
<td></td>
<td>c1)</td>
<td>Method 1</td>
<td>0.391</td>
</tr>
<tr>
<td></td>
<td>Method 2</td>
<td>0.466</td>
<td>.359</td>
<td>.855</td>
</tr>
<tr>
<td></td>
<td>Method 3</td>
<td>0.369</td>
<td>.278</td>
<td>.913</td>
</tr>
<tr>
<td></td>
<td>Method 4</td>
<td>0.375</td>
<td>.278</td>
<td>.909</td>
</tr>
<tr>
<td></td>
<td>Method 1</td>
<td>0.400</td>
<td>.311</td>
<td>.900</td>
</tr>
<tr>
<td></td>
<td>Method 2</td>
<td>0.444</td>
<td>.343</td>
<td>.867</td>
</tr>
<tr>
<td></td>
<td>Method 3</td>
<td>0.364</td>
<td>.278</td>
<td>.915</td>
</tr>
<tr>
<td></td>
<td>Method 4</td>
<td>0.373</td>
<td>.278</td>
<td>.909</td>
</tr>
<tr>
<td></td>
<td>Random</td>
<td>0.409</td>
<td>.319</td>
<td>.887</td>
</tr>
</tbody>
</table>

Note: KL = Kullback–Leibler.

In general, the discrepancies in the evaluation criteria observed under item Pool 1 were not large across the proposed item selection methods. This outcome may be a result of the characteristics of this item pool. The mean ability difference for the two latent classes was about 1 logit unit. The largest absolute difference in the item difficulty estimates was about 5.9 logit units. Although only a small percentage of items had such large differences, those items were the most frequently used. The separation delineated by the ability and item difficulty distributions between the two latent classes was relatively large.

The results for Pool 2 are presented in Table 2. Compared with Pool 1, the classification accuracy decreased dramatically. This may be due to the different distributions of item difficulty.
for Pool 2. In Pool 2, the differences between the item difficulty parameters were simulated to be 1 logit unit for every item. The largest estimated differences were 2.49 logit units, much smaller than that in Pool 1. The use of the KL, the reversed KL, or the adaptive KL information did not have much impact on the evaluation criteria. It was evident that Methods 3 and 4 performed better than Methods 1 and 2 in terms of classification accuracy. The differences between Methods 3 and 4 were minor. Method 2 performed the worst with the highest mean absolute bias in ability estimates and the lowest correlations between the true and the estimated ability parameters whereas Methods 3 and 4 performed slightly better than did Method 1. The random linear form produced evidently lower classification accuracy, although Method 2 had just slightly higher accuracy. The true and the estimated ability parameters were more highly correlated for Methods 1, 3, and 4 than that for the random linear test. These three methods produced smaller mean absolute bias in the ability parameter estimates than the linear test. However, Method 2 performed even worse than the random linear test in terms of these two evaluation criteria. A possible explanation for the poor performance of Method 2 might be the use of the boundaries around the current ability estimate in computing the KL information. These boundaries associated with the number of items administered may introduce noise in the estimation process given the total test length was 20. When 5 items were administered, the interval was 1.75. Even when 20 items were administered, the interval width was 0.88. These were not negligible values. If the estimates in the early stages of the test were far away from the true values, intervals with such large widths may lead to substantial deviation from the assumed convergence point.

Pool 3 differed from Pool 1 in that the difference in the mean ability for the two latent classes was smaller (different by 0.5 logit unit). The results for Pool 3 are summarized in Table 3. Overall, Methods 3 and 4 performed slightly better than Methods 1 and 2 with slightly higher classification accuracy, correlations between the true and the estimated ability parameters, and lower mean absolute bias across the three methods of computing the KL information. Method 2 performed the worst. Like that observed in Pool 1, all studied item selection methods yielded higher classification accuracy than did the random linear item selection method, but the estimation accuracy for the ability parameters was compromised.

### Table 3. Estimation of Ability Parameters and Classification Accuracy for Pool 3

<table>
<thead>
<tr>
<th>KL</th>
<th>Item selection algorithm</th>
<th>Absolute bias</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>M</td>
<td>SD</td>
<td>Correlation</td>
<td>Classification accuracy (%)</td>
<td></td>
</tr>
<tr>
<td>KL, $\text{KL}(c_1</td>
<td></td>
<td>c_2)$</td>
<td>Method 1</td>
<td>0.453</td>
<td>0.383</td>
<td>.823</td>
</tr>
<tr>
<td></td>
<td>Method 2</td>
<td>0.468</td>
<td>0.397</td>
<td>.811</td>
<td>99.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Method 3</td>
<td>0.447</td>
<td>0.336</td>
<td>.844</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Method 4</td>
<td>0.453</td>
<td>0.336</td>
<td>.831</td>
<td>99.7</td>
<td></td>
</tr>
<tr>
<td>Reversed KL, $\text{KL}(c_2</td>
<td></td>
<td>c_1)$</td>
<td>Method 1</td>
<td>0.469</td>
<td>0.401</td>
<td>.811</td>
</tr>
<tr>
<td></td>
<td>Method 2</td>
<td>0.486</td>
<td>0.426</td>
<td>.788</td>
<td>98.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Method 3</td>
<td>0.445</td>
<td>0.338</td>
<td>.844</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Method 4</td>
<td>0.460</td>
<td>0.338</td>
<td>.831</td>
<td>99.8</td>
<td></td>
</tr>
<tr>
<td>Adaptive KL, $\sum_i \text{KL}(c_i</td>
<td></td>
<td>\hat{c})$</td>
<td>Method 1</td>
<td>0.441</td>
<td>0.380</td>
<td>.834</td>
</tr>
<tr>
<td></td>
<td>Method 2</td>
<td>0.476</td>
<td>0.450</td>
<td>.790</td>
<td>97.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Method 3</td>
<td>0.437</td>
<td>0.330</td>
<td>.851</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Method 4</td>
<td>0.428</td>
<td>0.330</td>
<td>.856</td>
<td>99.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Random</td>
<td>0.381</td>
<td>0.294</td>
<td>.887</td>
<td>96.0</td>
<td></td>
</tr>
</tbody>
</table>

Note: KL = Kullback–Leibler.
Pool 4 was similar to Pool 3 in terms of ability distributions while similar to Pool 2 in terms of item difficulty distributions for the two latent classes. This combination of ability and item difficulty distribution resulted in the smallest separation across latent classes in Pool 4. Compared with Pools 2 and 3, Pool 4 yielded lower classification accuracy, slightly lower correlations between the true and the estimated ability parameters, and slightly higher absolute bias in the ability estimates. There is no clear advantage of using the KL, the reversed KL, or the adaptive KL information. Methods 1, 3, and 4 performed better than the random item selection method in terms of all three evaluation criteria. Method 2 evidently performed worse than any of the other three methods, even worse than the random linear test in terms of the three evaluation criteria (see Table 4).

Regarding item pool usage, items with larger difficulty differences were more frequently used as these items could provide more KL information. This can be illustrated using Methods 1 and 2 in computing the KL information. Figures 1 and 2 present the KL and the reversed KL information at three ability levels, −2, 0, and 2 for Method 1 in Pools 1 and 2, respectively. Similar patterns were observed for Pools 3 and 4. The x-axis represents the differences in the item difficulty estimates between the two latent classes (Class 1 - Class 2) for all 500 items by rank ordering from the smallest to the largest value. The y-axis represents the KL or the reversed KL information. At the theta point of −2 in Pool 1, the larger the absolute difference in item difficulty estimates between the two classes, the larger the KL or the reversed KL information was. The KL information was smaller than the reversed KL information at the lower end of the scale but larger at the upper end of the scale. One possible reason for this observation might be the following. The probability of a correct response for Latent Class 1 should be higher than that for Latent Class 2 as items are more difficult at the lower end of the scale for Class 2. For the same reason, the difference between $P$ (probability of a correct response) and $Q$ (probability of an incorrect response) should not be large for Class 1, but is much larger for Class 2 for the low-ability level of −2. Therefore, switching Latent Class 1 with Class 2 leads to larger reversed KL information. Using Item 267 as an example, item difficulty estimates for Classes 1 and 2 for this item were −2.766 and 2.880, $P = .683$, and $Q = .317$ for Class 1; and $P = .008$ and $Q = .992$ for Class 2 at theta point of −2. The KL information was 0.477 but the reversed KL information

### Table 4. Estimation of Ability Parameters and Classification Accuracy for Pool 4

<table>
<thead>
<tr>
<th>KL</th>
<th>Item selection algorithm</th>
<th>Absolute bias</th>
<th>Classification accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>KL, $KL(c_1</td>
<td></td>
<td>c_2)$</td>
<td>Method 1 0.379 0.296</td>
</tr>
<tr>
<td>Method 2 0.479 0.375</td>
<td>.816</td>
<td>76.1</td>
<td></td>
</tr>
<tr>
<td>Method 3 0.367 0.280</td>
<td>.899</td>
<td>83.7</td>
<td></td>
</tr>
<tr>
<td>Method 4 0.373 0.280</td>
<td>.893</td>
<td>83.2</td>
<td></td>
</tr>
<tr>
<td>Reversed KL, $KL(c_2</td>
<td></td>
<td>c_1)$</td>
<td>Method 1 0.390 0.301</td>
</tr>
<tr>
<td>Method 2 0.490 0.381</td>
<td>.807</td>
<td>76.3</td>
<td></td>
</tr>
<tr>
<td>Method 3 0.368 0.281</td>
<td>.898</td>
<td>83.4</td>
<td></td>
</tr>
<tr>
<td>Method 4 0.388 0.281</td>
<td>.886</td>
<td>82.7</td>
<td></td>
</tr>
<tr>
<td>Adaptive KL, $\sum_{i} KL(c_i</td>
<td></td>
<td>\hat{c})$</td>
<td>Method 1 0.376 0.288</td>
</tr>
<tr>
<td>Method 2 0.476 0.368</td>
<td>.822</td>
<td>76.2</td>
<td></td>
</tr>
<tr>
<td>Method 3 0.363 0.274</td>
<td>.902</td>
<td>82.5</td>
<td></td>
</tr>
<tr>
<td>Method 4 0.375 0.274</td>
<td>.894</td>
<td>82.9</td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td>0.393 0.294</td>
<td>.881</td>
<td>79.7</td>
</tr>
</tbody>
</table>

Note: KL = Kullback–Leibler.
was 1.179 at theta = 2. Following the same logic, the KL information should be larger than the reversed KL information at the upper end of the scale. At the theta level of 0, the KL information was about the same as the reversed KL information. This is due to the fact that the discrepancy in the differences between $P$ and $Q$ for Classes 1 and 2 decreased. Again using Item 267 as an example, $P = .941$ and $Q = .059$ for Class 1, and $P = .053$ and $Q = .947$ for Class 2 at theta = 0. The KL information was 1.074, whereas the reversed KL information was 1.103 at theta = 0. This reversed pattern of $P$ and $Q$ for Classes 1 and 2 leads to similar KL and reversed KL information. At the theta level of 2, the KL information was larger than the reversed KL.
information at the lower end of the scale, but smaller at the upper end of the scale. Closely related to the KL and the reversed KL information distribution, about 25 items had very high frequency of usage with a frequency of more than 1,000 times to even 10,000 times.

For Pool 2, the differences in item difficulty parameters between the two classes were simulated to be 1. With smaller item difficulty discrepancy, the difference between the KL and the reversed KL information provided by the same item was much smaller than that observed in Pool 1. However, the amount of information was close to that observed for the similar range of item difficulty differences found in Pool 1. Across items, larger information was associated with larger absolute item difficulty differences, but the amount of information provided by items with large difficulty differences varied more than that in Pool 1 (see Figure 2). About 36 items had very high frequency of usage with a frequency of more than 1,000 times to around 8,500 times. This difference in the number of items frequently used between Pools 1 and 2 is due to the amount of information provided by individual items. A majority of the items in Pool 1 provided much higher information than those in Pool 2 which can be seen in Figures 1 and 2.

Item usage was further examined under Method 2 which set up an interval around the current ability estimate. The lower bound was paired with Class 1, which had a lower mean ability, whereas the upper bound was paired with Class 2, which had a higher mean ability. For Pool 1, similar patterns to those found under Method 1 were observed, but the differences between Methods 1 and 2 were distinct in terms of the magnitude of the information and the interval with larger information (see Figure 3). The items with large difficulty differences at the high end of the scale provided much more information than those at the lower end of the scale. This is due to the pairing of items and the ability bounds. At the lower end of the scale, Class 1 with easier items was paired with the lower ability bound, whereas Class 2 with more difficult items was paired with the higher ability bound. This leads to smaller differences between $P$ and $Q$, thus, smaller KL and reversed KL information. At the higher end of the scale, Class 1 with more difficult items was paired with the lower ability bound, whereas Class 2 with easier items was paired with the higher ability bound. This leads to much larger difference between $P$ and $Q$, thus larger KL and reversed KL information. Therefore, more items which were more difficult for Class 1 and easier for Class 2 were used under Method 2. This is different from that observed in Method 1 where items at the two ends of the scale with large absolute item difficulty differences were equally likely to be used. At the theta level of $-2$, the KL information was smaller than the reversed KL information at the lower end of the scale, but larger at the upper end of the scale. There was not much difference at theta = 0. At the theta level of 2, the KL information was larger than the reversed KL information at the lower end of the scale, but smaller at the upper end of the scale. The magnitude of the KL and the reversed KL information for Method 2 was smaller than that observed under Method 1 for Pool 1 at the lower end of the scale but larger at the upper end of the scale.

For Pool 2, the item difficulty differences ranged from around $-2$ to $-2.5$. For this given interval, the information at the upper end of the scale was much larger than that at the lower end of the scale, which essentially approached zero as shown in Figure 4. The pattern presented in Figure 4 is consistent with that presented in Figure 3 for this similar range of item difficulty differences. Across items, larger information was associated with larger item difficulty differences at the upper end of the scale, but the amount of information provided by items with large difficulty differences varied more than that in Pool 1 (see Figure 4). Item pool usage was similar to that found in Method 1. For Pools 1 and 2, the maximum information provided by Method 2 was more than that provided by Method 1. This is due to the use of the upper and lower bounds of the interval set around the ability estimate that enhance the discrimination power between the two latent classes. As Methods 3 and 4 require the current estimates of class-specific ability,
it is difficult to effectively present the KL information distribution. Thus, only the KL and the reversed KL information are presented for Methods 1 and 2 for illustration.

In general, about 25 items had very high frequency of usage with a frequency of more than 1,000 times to even 10,000 times for Pools 1 and 3 across methods. About 36 items had very high frequency of usage with a frequency of more than 1,000 times to around 8,500 times for Pools 2 and 4 across methods.

The distribution of the converged posterior classification decisions as a function of item sequence (5-20) in the CAT administration is summarized in Table 5 for Pool 1. When the

Figure 2. KL and RKL information at three theta points for Method 1 in Pool 2
Note: KL = Kullback–Leibler; RKL = reversed KL.
classification decision for an examinee no longer changes after the administration of an item, the classification decision is defined to stabilize at that item. For all methods compared, the classification became stabilized or converged for more than 70% of the examinees after the administration of the first five items in Pool 1. After the administration of Item 6, Method 1 (including all three alternatives under this method) showed higher percentage of confident classification decisions (more than 90%) than did Method 4 (more than 85%). The confidence in classification was roughly monotonically decreasing over the course of the test. Similar patterns have been observed for Pool 3.

**Figure 3.** KL and RKL information at three theta points for Method 2 in Pool 1
Note: KL = Kullback–Leibler; RKL = reversed KL.

---

Jiao et al.
Table 6 summarizes the same information for Pool 2. The number of examinees whose classification converged at Item 5 was smaller than that for Pool 1. This is due to less KL information provided by Pool 2 as shown in the figures. Thus, more items need to be administered before classification decisions stabilize. All alternatives under Method 2 required fewer items.

Figure 4. KL and RKL information at three theta points for Method 2 in Pool 2
Note: KL = Kullback–Leibler; RKL = reversed KL.

Table 6 summarizes the same information for Pool 2. The number of examinees whose classification converged at Item 5 was smaller than that for Pool 1. This is due to less KL information provided by Pool 2 as shown in the figures. Thus, more items need to be administered before classification decisions stabilize. All alternatives under Method 2 required fewer items.
<table>
<thead>
<tr>
<th>Item</th>
<th>Method 1</th>
<th>Method 2</th>
<th>Method 3</th>
<th>Method 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>7,528</td>
<td>7,629</td>
<td>7,391</td>
<td>7,380</td>
</tr>
<tr>
<td>6</td>
<td>1,586</td>
<td>1,413</td>
<td>1,354</td>
<td>1,424</td>
</tr>
<tr>
<td>7</td>
<td>510</td>
<td>511</td>
<td>551</td>
<td>479</td>
</tr>
<tr>
<td>8</td>
<td>155</td>
<td>234</td>
<td>295</td>
<td>312</td>
</tr>
<tr>
<td>9</td>
<td>82</td>
<td>93</td>
<td>133</td>
<td>160</td>
</tr>
<tr>
<td>10</td>
<td>44</td>
<td>44</td>
<td>97</td>
<td>101</td>
</tr>
<tr>
<td>11</td>
<td>23</td>
<td>23</td>
<td>55</td>
<td>44</td>
</tr>
<tr>
<td>12</td>
<td>16</td>
<td>13</td>
<td>20</td>
<td>35</td>
</tr>
<tr>
<td>13</td>
<td>14</td>
<td>12</td>
<td>30</td>
<td>24</td>
</tr>
<tr>
<td>14</td>
<td>8</td>
<td>5</td>
<td>12</td>
<td>14</td>
</tr>
<tr>
<td>15</td>
<td>5</td>
<td>3</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>16</td>
<td>4</td>
<td>4</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>17</td>
<td>8</td>
<td>4</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>19</td>
<td>5</td>
<td>3</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>20</td>
<td>6</td>
<td>8</td>
<td>23</td>
<td>1</td>
</tr>
</tbody>
</table>

Note: KL = Kullback-Leibler.
Table 6. The Number of Examinees Whose Classification Converged at a Specific Item for Pool 2

<table>
<thead>
<tr>
<th>Item</th>
<th>Method 1</th>
<th>Method 2</th>
<th>Method 3</th>
<th>Method 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>4,490</td>
<td>4,569</td>
<td>4,555</td>
<td>5,589</td>
</tr>
<tr>
<td>6</td>
<td>1,150</td>
<td>1,105</td>
<td>1,098</td>
<td>1,696</td>
</tr>
<tr>
<td>7</td>
<td>454</td>
<td>443</td>
<td>566</td>
<td>838</td>
</tr>
<tr>
<td>8</td>
<td>421</td>
<td>382</td>
<td>370</td>
<td>603</td>
</tr>
<tr>
<td>9</td>
<td>487</td>
<td>371</td>
<td>336</td>
<td>404</td>
</tr>
<tr>
<td>10</td>
<td>332</td>
<td>341</td>
<td>328</td>
<td>296</td>
</tr>
<tr>
<td>11</td>
<td>273</td>
<td>260</td>
<td>289</td>
<td>223</td>
</tr>
<tr>
<td>12</td>
<td>250</td>
<td>203</td>
<td>227</td>
<td>138</td>
</tr>
<tr>
<td>13</td>
<td>256</td>
<td>276</td>
<td>303</td>
<td>93</td>
</tr>
<tr>
<td>14</td>
<td>257</td>
<td>260</td>
<td>246</td>
<td>50</td>
</tr>
<tr>
<td>15</td>
<td>279</td>
<td>276</td>
<td>257</td>
<td>26</td>
</tr>
<tr>
<td>16</td>
<td>223</td>
<td>285</td>
<td>228</td>
<td>17</td>
</tr>
<tr>
<td>17</td>
<td>271</td>
<td>282</td>
<td>242</td>
<td>9</td>
</tr>
<tr>
<td>18</td>
<td>257</td>
<td>263</td>
<td>244</td>
<td>9</td>
</tr>
<tr>
<td>19</td>
<td>263</td>
<td>339</td>
<td>306</td>
<td>5</td>
</tr>
<tr>
<td>20</td>
<td>337</td>
<td>345</td>
<td>405</td>
<td>4</td>
</tr>
</tbody>
</table>

Note: KL = Kullback–Leibler.
to produce stable classification decisions for a majority of the examinees. The other three methods showed similar patterns. The classification convergence speed displayed a monotonically decreasing pattern in Method 2 whereas other methods generally flattened at Item 11. This again can be explained by the information presented in Figures 2 and 4. Method 2 produced much higher KL (including the KL and the reversed KL) information than did Method 1 for Pool 2.

In summary, when item separation was large as that observed in Pools 1 and 3, there was essentially not much difference in classification accuracy and estimation accuracy in ability parameters across methods. The magnitude of separation in ability distributions of the two latent classes did not have much impact on the evaluation criteria. However, for the smaller item separation as that observed in Pools 2 and 4, there was a dramatic reduction in the classification accuracy (about 15% decrease) although the impact on the other two evaluation criteria was relatively smaller. The use of the KL, the reversed KL, or the adaptive KL information did not have much impact on the results. More items were used with high frequency when item difficulty differences were smaller. The classification stabilization point out of the administered 20 items depended on the item pool features as well as the proposed methods.

Summary and Discussions

This study proposed four indices adapted from the KL information to select items in the mixture Rasch model–based CAT administration. The four methods focused on latent class identification and were illustrated by classifying examinees into one of two latent classes given an item pool calibrated with the mixture Rasch model. Two of these methods were developed in line with the practice of estimating single latent ability across latent classes, whereas the remaining two were developed in line with the practice of estimating class-specific latent ability. The four methods for item selection based on the KL information were compared with the reversed and the adaptive KL information.

In general, the simulation studies supported the use of Method 1 over Method 2 given single latent ability estimated across all latent classes. Methods 3 and 4 were viable item selection methods when class-specific latent ability parameters were estimated. Overall, to achieve higher classification accuracy in the mixture Rasch model–based CAT, class-specific ability should be estimated for each latent class, and Method 3 might be recommended as it maximizes information from all possible sources and matches each source of information accordingly.

When item difficulty separation was large as that simulated in Pools 1 and 3, all four methods sacrificed accuracy in ability estimation in exchange for enhanced classification accuracy over random linear item selection. However, all methods except Method 2 lead to more accurate classification decisions and ability estimation than random item selection in Pools 2 and 4 where item difficulty separation was smaller.

The four KL information indices are adapted under the assumption that the latent ability is unidimensional across latent classes. When different latent traits are measured across latent classes, that is, in the presence of between-class multidimensionality, class-specific latent traits need to be estimated. Thus, only Method 3 would be appropriate for item selection and only the posterior probability of each latent class based on its own class-specific ability estimate is proper for use in estimating the latent class membership.

No theoretical rationale is provided in literature related to the positioning of the probability functions in computing the KL information. This study provided empirical evidence indicating that the KL and the reversed KL information differed given the same item and ability parameters. However, the use of the KL or the reversed KL information had little impact on the outcome variables; that is, overall classification accuracy and ability estimation accuracy under the
simulated study conditions. This might be due to the relative symmetric distribution of item difficulty differences in each simulated item pool. The methods can be replicated to other item pools with different characteristics and other mixture IRT models.

In addition to the relative performance of the proposed methods for item selection, the findings of this study may shed some lights on item pool construction for the mixture Rasch model–based CAT administration. In general, to maximize classification accuracy and efficiency, items with very disparate item difficulty parameters should be selected to the item pool. Items that are very easy for one class but very difficult for the other class will provide much more KL information.

This study found that items with larger difficulty differences between the two latent classes were more frequently selected for administration. It is worthy of future exploration to investigate whether the effect of item discrimination parameter in item selection is similar to that found in the conventional CAT based on the unidimensional non-Rasch IRT models (Chang & Ying, 1999). It would also be interesting to explore the generalization of the proposed methods to the polytomous mixture Rasch model (Rost, 1991; von Davier & Rost, 1995) when polytomous items are often used in surveys in different fields to identify latent classes.

Under Method 2, the upper bound of the latent ability estimate was paired with the latent class with a higher mean ability, whereas the lower bound was paired with the latent class with a lower mean ability. This study further examined the KL and the reversed KL information distribution for Method 2 in Pools 1 and 2 when the pairing was reversed. In general, the KL and the reversed KL information distribution patterns were flipped compared with those based on the original pairing. For Pool 1, the magnitude of the KL information was not affected. For Pool 2, however, the magnitude of the KL and the reversed KL information decreased. This indicates that the original pairing provided more or at least the same amount of information.

Several limitations of this study need to be addressed in future explorations. This study only investigated item selection without consideration of other “real-world” constraints, such as item exposure control and content balance. This may be an important issue because substantial overlap of item usage across examinees is found. For test security, item exposure should be controlled in real-world implementation of a CAT. A more constrained version of the simulated CAT should be investigated in future studies. Furthermore, the CAT was terminated using a fixed length criterion; every simulated examinee adaptively responded to 20 items. Measurement precision was different for different ability levels. A variable test length but with the same measurement precision of ability should be explored in the future. As Chang and Ying (1996) have suggested, δ related to Method 2 is also worthy of study in such applications. Two values of the c constant: 0.95 and 1.96 in the computation of δ were explored. Only the results related to the latter were reported in this study because the results from these two values were very similar. However, it is still worthy of future investigation to compare other values.

The present study focuses on classification. Further exploration is needed centering on ability estimation accuracy or on classification and ability estimation accuracy. Another possible study would be to compare item selection based on the Fisher information with the item selection methods proposed in this study, and investigate their respective impact on classification and ability estimation accuracy.

As one reviewer has pointed out, the impact of the initial item selection and examinee latent class assignment on the final classification also warrants attention. Item difficulties for the first five items randomly selected could vary greatly across examinees conditional on the latent class membership. This may induce potential threats to test fairness. The current initial assignment of latent class membership only works under the assumption of ordered latent classes. If latent classes are not ordered, new methods for initial class assignment should be explored. Future
research may explore other alternatives for initial class assignment and the consequences on the 
final classification decisions when latent classes are ordered or unordered in terms of ability.

In summary, this study generally supports the use of three proposed item selection methods: 
Methods 1, 3, and 4 in the mixture Rasch model–based CAT administration, especially for latent 
class identification. The KL, the reversed KL, and the adaptive KL information, in general, did 
not produce very different results. Although it is not known whether any operational CAT is cur-
rently implemented based on the mixture IRT models, the potential advantages of using the 
approaches explored in the present study should not be ignored if the purpose of a test is to iden-
tify the latent class membership of examinees.

Acknowledgments
The authors would like to thank Dr. Robert Mislevy for constructive feedback to an earlier version of the 
paper presented at the 2011 annual meeting of the National Council on Measurement in Education. They 
particularly thank the editor, Dr. Hua-Hua Chang, and the three anonymous reviewers for their suggestions, 
which greatly improved the manuscript.

Declaration of Conflicting Interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or pub-
lication of this article.

Funding
The author(s) received no financial support for the research, authorship, and/or publication of this article.

References
Dekker.
speededness: Application of a mixture Rasch model with ordinal constraints. Journal of Educational 
Measurement, 39, 331-348.
distributions in the hybrid model for test speededness: A comparison of marginal maximum likelihood 
estimation and Markov chain Monte Carlo estimation. Paper presented at the annual meeting of the 
Assessment, 12, 283-301.
Psychological Measurement, 23, 211-222.


