Graduate
Management
Admission
Council®

Item Selection and Hypothesis Testing for the Adaptive Measurement of Change

Matthew Finkelman
Tufts University School of Dental Medicine
David J. Weiss
University of Minnesota
Gyenam Kim-Kang
Korea Nazarene University

Presented at the Item Selection Paper Session, June 2, 2009

Abstract

Assessing individual change is an important topic in both psychological and educational measurement. An adaptive measurement of change (AMC) method had previously been shown to exhibit greater efficiency in detecting change than conventional non-adaptive methods. However, little work had been done to compare different procedures within the AMC framework. This study introduced a new item selection criterion and two new test statistics for detecting change with AMC that were specifically designed for the paradigm of hypothesis testing. In two simulation sets, the new methods for detecting significant change improved upon existing procedures by demonstrating better adherence to Type I error rates and substantially better power for detecting relatively small change, while substantially reducing test lengths when the second adaptive test was variable length.

Acknowledgment

Presentation of this paper at the 2009 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®. A revised version of this paper has been published in *Applied Psychological Measurement* (2010), Volume 34, Number 4, pp. 238–254.

Copyright © 2009 by the Authors

Citation

Finkelman, M., Weiss, D. J., & Kim-Kang, G. (2009). Item Selection and Hypothesis Testing for the Adaptive Measurement of Change. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* Retrieved [date] from www.psych.umn.edu/psylabs/CATCentral/

Author Contact

**Matthew Finkelman, Tufts University School of Dental Medicine,
Department of Research Administration, 75 Kneeland Street, Room 105,
Boston, MA 02111, U.S.A. Email: mattstat2000@yahoo.com**

Item Selection and Hypothesis Testing for the Adaptive Measurement of Change

In psychological testing, the assessment of individual change is often critical when tracking the trajectory of a patient. For instance, a practitioner might measure a patient's initial level of severity along some domain (such as depression, phobia, or headache impact) using a questionnaire or inventory. Once the patient has undergone treatment, another questionnaire is administered for purposes of comparison. A statistical test might then be used to determine whether significant improvement or decline (or neither) has occurred since the initial measurement.

Assessing individual change is also important in the field of educational testing. Teachers often seek to ascertain whether a student has advanced or regressed between examinations. Advancement might be made through teaching or practice of the material; regression might occur if the student forgets material that had previously been mastered. Because the diagnosis of change is again made on the basis of responses to test items, the same psychometric and statistical issues are common to measuring individual change in both the educational and psychological realms.

Measuring individual change has been controversial in the psychometric literature (Bereiter, 1963; Cronbach & Furby, 1970; Embretson, 1995). Since Cronbach and Furby's call for a moratorium on attempts to measure individual change, based on their evaluation of classical test theory methods for measuring change, little progress has been made on this important problem. However, Kim-Kang and Weiss (2007, 2008) combined the modern technologies of item response theory (IRT) and computerized adaptive testing (CAT) in a procedure, originally proposed by Weiss and Kingsbury (1984), that they called adaptive measurement of change (AMC). They then compared four procedures for measuring individual change using monte-carlo simulation. The four procedures were the simple difference score (Burr & Nesselroade, 1990; McDonald, 1999), residual change score (Manning & DuBois, 1962; Willett, 1997), a difference score based on IRT (Kim-Kang & Weiss, 2007, 2008), and AMC. Kim-Kang and Weiss concluded that AMC better captured actual levels of change and greatly improved the efficiency of detecting change compared to the other three methods. However, because their goal was not to compare different procedures *within* AMC, their study of this approach was limited to one item selection criterion and one method of hypothesis testing. The maximum information criterion of item selection that they used is suitable when the practitioner's goal is to *estimate* change, but might be improved upon when the goal is a powerful *hypothesis test* of change. Additionally, although their use of confidence intervals for hypothesis testing was superior to non-adaptive methods for detecting change, the resulting power was low under a number of circumstances investigated. Thus, new statistical tests need to be developed and investigated to improve AMC's power to detect significant change.

The purpose of this study was to extend the work of Kim-Kang and Weiss (2007, 2008) by introducing a new item selection method and adapting two hypothesis testing methods to AMC. The new item selection criterion is specifically designed for the paradigm of hypothesis testing, as opposed to that of estimation. Therefore, it has the potential to enhance the statistical power when testing whether significant change has occurred in a patient or student. The new hypothesis tests were adapted from standard hypothesis testing methods and applied to detecting change within AMC. These item selection and hypothesis testing methods were compared to existing procedures in simulation.

Item Response Theory and AMC

Item response theory (IRT) is a well-known psychometric tool for relating an examinee's latent trait of interest (e.g., severity, ability, achievement) to his/her responses to a test or other measuring instrument (Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980). The three-parameter logistic model (3PLM) is a popular choice in many testing applications when responses are scored dichotomously as correct or incorrect (or "keyed/not keyed"), as was assumed in the present research. Let $u_j = 1$ if item j is answered correctly, and $u_j = 0$ otherwise. Under the 3PLM

$$P\left(u_j = 1 | \theta\right) \equiv p_j\left(\theta\right) = c_j + \frac{1 - c_j}{1 + \exp\left[-1.7 a_j \left(\theta - b_j\right)\right]} \qquad (1)$$

(Birnbaum, 1968; Lord, 1980), where a_j is the item's slope (or discrimination), b_j is its difficulty (or location), c_j is its pseudo-guessing parameter, and θ represents the latent trait of interest.

One of the advantages of IRT is that different examinees (or the same examinee at different time points) can be placed on the same scale even when different items are used across measurements. This feature facilitates the use of CAT (van der Linden & Glas, 2000; Wainer, 2000; Weiss, 1983; Weiss & Kingsbury, 1984). In CAT, an instrument is tailored to examinees at the individual level, typically selecting the most informative items for each examinee based on his/her previous answers. CAT can also determine when enough information has accrued and no more items need to be presented to each examinee. In this way, measurement can be made both more precise and more efficient, as examinees receive items congruent with their latent trait and the test can be ended when a sufficient number of items has been given.

In AMC, there are two (or more) measurements taken: one at the first time point of interest (Time 1) and the other at a later time point of interest (Time 2). Let θ_1 denote the value of an arbitrary examinee's true latent trait at Time 1, and θ_2 the corresponding value at Time 2. To determine whether meaningful change has occurred for an examinee (e.g., learning, clinical improvement, decline), the practitioner's goal is to determine whether θ_2 is different from θ_1 using hypothesis testing. Since change might occur in either direction, a two-sided test is appropriate, with hypotheses $H_0: \ \theta_1 = \theta_2$ and $H_1: \ \theta_1 \neq \theta_2$. Either fixed-length testing or variable-length testing can be employed in the procedure. In the former case, the test lengths at the two time points are pre-specified; in the latter case, they are random variables. Hereafter, the test length at Time 1 will be denoted *K* and the test length at Time 2 will be denoted *L*.

In order to perform item selection in AMC, candidate items must be designated for selection at each time point. There are three possibilities: (1) use two separate item banks, one for each time point, with the banks linked onto a common scale; (2) use the same item bank at each time point but require for a given examinee that all items administered at Time 1 be ineligible for that examinee at Time 2; or (3) use the same item bank at each time point and allow all items to be eligible at Time 2 for all examinees. In educational testing, option 3 is often inappropriate because students might remember the items from time point to time point. However, in psychological assessment, it might be perfectly reasonable for an examinee to receive an item at both time points. As in Kim-Kang and Weiss (2007, 2008), option 1 or 3 was assumed in this study. All hypothesis testing methods examined here are still applicable when option 2 is used.

Item Selection and Hypothesis Testing Methods

Previous Work

Item selection. Most CAT item selection methods require that an *interim* θ *estimate* be made after each item response. Suppose first that the practitioner is interested only in an examinee's current θ level, rather than the measurement of change. It is then typically assumed that k items have been presented, eliciting a response vector $\mathbf{u}_k = (u_1, \ldots, u_k)$, and that the stopping rule has not yet been invoked (so the CAT will administer at least one more item). Under the 3PLM, and making the usual IRT assumption of local independence, the likelihood function is defined as

$$L(\theta, \mathbf{u}_k) = \prod_{j=1}^{k} p_j(\theta)^{u_j} \left[1 - p_j(\theta)\right]^{1-u_j} \qquad (2)$$

(Lord, 1980). The value of θ maximizing Equation 2 is called the *maximum likelihood estimate* (MLE) of θ after k items; this will be denoted $\hat{\theta}_k$. The MLE is the value that satisfies

$$\sum_{j=1}^{k} \left[u_j - p_j(\theta)\right] \frac{p_j'(\theta)}{p_j(\theta)\left[1 - p_j(\theta)\right]} = 0 \qquad (3)$$

(Lord, 1980), where $p_j'(\theta)$ is the first derivative of the logistic curve evaluated at θ. This value can be considered as the value of θ that is most consistent with the response vector $\mathbf{u}_k$. The MLE is often used because it does not require the use of a subjective prior distribution as do Bayesian approaches such as the expected *a posteriori* (EAP) and maximum *a posteriori* (MAP) estimators (Embretson & Reise, 2000).

Once the interim θ estimate has been made, this estimate can be used to select item $k+1$. A traditional procedure is to choose the item that maximizes the Fisher information at $\hat{\theta}_k$, among all items that have not yet been administered to the examinee (for example, see Lord, 1980). The Fisher information at $\hat{\theta}_k$ for the 3PLM is computed as (Lord, 1980)

$$I_j(\hat{\theta}_k) = \frac{1.7^2 a_j^2 (1 - c_j)}{\left\{c_j + \exp\left[1.7 a_j (\hat{\theta}_k - b_j)\right]\right\}\left\{1 + \exp\left[-1.7 a_j (\hat{\theta}_k - b_j)\right]\right\}^2}. \qquad (4)$$

When the MLE does not exist, an item is typically selected to maximize the Fisher information at an arbitrary value of θ. For instance, the first item might be chosen to maximize Equation 4 at $\theta = 0$; when all previously administered k items have been answered correctly, item $k+1$ might be chosen to maximize Equation 4 at a high value of θ; when all previously administered k items have been answered incorrectly, item $k+1$ might be chosen to maximize Equation 4 at a low value of θ.

In AMC, item selection must be conducted for the same examinee at two (or more) different time points (and possibly two different levels of the underlying latent trait). For this case, Kim-Kang and Weiss (2007, 2008) treated the two time points as distinct measurements and used Equation 4 separately for each measurement. That is, at Time 1, Kim-Kang and Weiss selected item $k+1$ to maximize $I_j(\hat{\theta}_1^k)$, where $\hat{\theta}_1^k$ denotes the Time 1 MLE after k items. Similarly, at Time 2, they selected item $l+1$ to maximize $I_j(\hat{\theta}_2^l)$, where $\hat{\theta}_2^l$ is the Time 2 MLE after l items. As a result, responses from Time 1 were not taken into account when selecting items at Time 2. The only exception occurred when selecting the first item at Time 2; in that situation, Kim-Kang and Weiss maximized Fisher information at the final Time 1 MLE. For notational simplicity, the final Time 1 MLE is hereafter denoted $\hat{\theta}_1$ and the final Time 2 MLE is denoted $\hat{\theta}_2$; that is, the subscripts k and l are suppressed for final θ estimates.

Hypothesis testing. Once both the Time 1 and Time 2 measurements have concluded, AMC uses a statistical significance test to determine whether change has occurred (that is, whether $\theta_1 \neq \theta_2$). Kim-Kang and Weiss (2007, 2008) calculated separate confidence intervals for θ_1 and θ_2, then identified significant change based on whether the two confidence intervals overlapped. A $(1-\alpha)\times 100\%$ confidence interval for θ_1 is given by

$$\hat{\theta}_1 \pm z_{1-\alpha/2}\, SE\left(\hat{\theta}_1\right), \tag{5}$$

where $z_{1-\alpha/2}$ is the appropriate quantile of the standard normal distribution and $SE\left(\hat{\theta}_1\right)$ is the standard error of measurement evaluated at $\hat{\theta}_1$. Kim-Kang and Weiss computed the standard error at $\hat{\theta}_1$ as

$$SE\left(\hat{\theta}_1\right) = \sqrt{-\left[I\left(\hat{\theta}_1\right)\right]^{-1}}, \tag{6}$$

where $I\left(\hat{\theta}_1\right)$ is the observed information at $\hat{\theta}_1$:

$$I\left(\hat{\theta}_1\right) = \sum_{j=1}^{K_1} 1.7^2 a_j^2 \left(\frac{p_j\left(\hat{\theta}_1\right) - c_j}{1-c_j}\right)\left(\frac{1-p_j\left(\hat{\theta}_1\right)}{1-c_j}\right)\left(\frac{u_j c_j - p_j\left(\hat{\theta}_1\right)^2}{p_j\left(\hat{\theta}_1\right)^2}\right) \tag{7}$$

(Baker & Kim, 2004). The confidence interval for θ_2 is analogous. Once confidence intervals for θ_1 and θ_2 have been calculated, statistically significant change is identified if and only if these intervals do not overlap—that is, if the lower bound for θ_2 is higher than the upper bound for θ_1, or vice-versa.

A New Item Selection Method for Time 2 in AMC

The item selection method of Kim-Kang and Weiss (2007, 2008) is a reasonable approach for AMC. After all, a rule maximizing information at the Time 1 MLE is designed to yield a precise estimate of θ_1; maximizing information at the Time 2 MLE is designed to yield a precise estimate of θ_2. As noted previously, however, such a procedure does not consider the information accrued at Time 1 when selecting items at Time 2. In the current application of AMC, where the goal is not estimation but rather a hypothesis test of change between time points, it might be fruitful to take such information into account. That is, if items at Time 2 are selected with the explicit purpose of differentiating between θ_1 and θ_2, the power to detect true change might be enhanced. The new item selection procedure for Time 2 incorporates the examinee's answers at Time 1. This method changes only the item selection criterion at Time 2; at Time 1, maximum Fisher information at $\hat{\theta}_1^k$ is still employed.

The method involves the notion of Kullback-Leibler information, which is prevalent in the discipline of Information Theory (Cover & Thomas, 1991) and was first applied to CAT by Chang and Ying (1996). Let θ' and θ'' be two candidate θ values at the time point of interest (here, Time 2). The Kullback-Leibler information of item j for distinguishing these values is equal to

$$K_j(\theta', \theta'') = E_{\theta'}\left[\ln \frac{L(\theta'; u_j)}{L(\theta''; u_j)}\right], \tag{8}$$

where $E_{\theta'}[X]$ denotes the expectation of X under θ'. A large value of $K_j(\theta', \theta'')$ indicates that item j is useful in differentiating between θ' and θ'' when θ' is the true state of nature. Hence, high power in AMC will be achieved if θ' is the best estimate of θ_2 under the assumption that H_1 is true, and θ'' is the best estimate of θ_2 under the assumption that H_0 is true.

Assume first that change has indeed occurred between Time 1 and Time 2, i.e., that H_1 is correct. By this assumption, all responses at Time 1 were elicited from a level of θ that is not the true state of nature at Time 2. If no *a priori* information exists about the anticipated magnitude of change between time points, then responses from Time 1 do not convey information about the true value of θ_2. Hence, after l items, the MLE of θ_2 under H_1 is simply $\hat{\theta}_2^l$, assuming that this value differs from $\hat{\theta}_1^k$.

By contrast, if H_0 is true, then both the Time 1 and Time 2 observations arose from the same value of θ. Therefore, the responses from both time points can be combined or "pooled" to obtain an overall estimate of this single θ value. Let $\boldsymbol{u}_{K+\ell}$ denote the vector containing all $K+l$ observations at the two time points (K items having been administered at Time 1 and l items having been administered thus far at Time 2). Then the MLE under H_0 (i.e., the "pooled MLE") is the value of θ maximizing

$$L(\theta;\, \boldsymbol{u}_{K+l}) = \prod_{j=1}^{K+l} p_j(\theta)^{u_j}\left[1 - p_j(\theta)\right]^{1-u_j}. \tag{9}$$

This maximizing value, which will be denoted $\hat{\theta}_{pool}^{K+l}$, is that which satisfies Equation 3 for the vector $\boldsymbol{u}_{K+\ell}$.

Based on the logic outlined above, the new item selection method sets $\theta' = \hat{\theta}_2^l$ and $\theta'' = \hat{\theta}_{pool}^{K+l}$, i.e., item $l+1$ at Time 2 is chosen to maximize $K_j(\hat{\theta}_2^l, \hat{\theta}_{pool}^{K+l})$. This method is reminiscent of that of Eggen (1999), who also maximized the Kullback-Leibler information at two distinct θ values in the context of computerized classification testing. However, in Eggen's case these two values were fixed, whereas $\hat{\theta}_2^l$ and $\hat{\theta}_{pool}^{K+l}$ both change over time and thus allow for adaptation. In the unlikely event that $\hat{\theta}_2^l = \hat{\theta}_{pool}^{K+l}$, the maximum $K_j(\hat{\theta}_2^l, \hat{\theta}_{pool}^{K+l})$ criterion cannot be used, since there is no Kullback-Leibler information between a value and itself. In this case, or if no items have yet been administered at Time 2 (so that $l = 0$), Fisher information at $\hat{\theta}_1$ can be utilized as a substitute item selection method.

New Hypothesis Testing Methods for AMC

As above, the null hypothesis is no change between time points (i.e., $\theta_1 = \theta_2$) and the alternative hypothesis is that change has occurred in either direction $(\theta_1 \neq \theta_2)$. Fixed-length tests are initially assumed at both time points; the analysis is then extended to variable-length testing. As with the notations $\hat{\theta}_1$ and $\hat{\theta}_2$, the staging subscript will be dropped for the final pooled estimate, i.e., $\hat{\theta}_{pool}$ will denote the pooled estimate after both tests have been completed.

The Z-test. This approach tests the null hypothesis using the standardized difference in MLE values. This test statistic is defined as

$$|Z| = \frac{|\hat{\theta}_2 - \hat{\theta}_1|}{\sqrt{\dfrac{1}{\sum_{j=1}^{K} I_j\left(\hat{\theta}_{pool}\right)} + \dfrac{1}{\sum_{j=1}^{L} I_j\left(\hat{\theta}_{pool}\right)}}}, \tag{10}$$

where $\sum_{j=1}^{K} I_j\left(\hat{\theta}_{pool}\right)$ is the sum of the Fisher information values at $\hat{\theta}_{pool}$ for the Time 1 items, and $\sum_{j=1}^{L} I_j\left(\hat{\theta}_{pool}\right)$ is the analogous quantity for Time 2. In a long test, the inverse of the summed item information can be used to approximate the variance of the MLE (Chang & Stout, 1993); here, the information is evaluated at $\hat{\theta}_{pool}$ because this value is the most plausible θ level under the null hypothesis. By the additivity of variances under local independence, the denominator in Equation 10 can be considered the standard error of $\hat{\theta}_2 - \hat{\theta}_1$ under the null. Moreover, each MLE is asymptotically normal under mild regularity conditions (Chang & Stout, 1993), so their difference is also asymptotically normal. The absolute value is taken because the alternative hypothesis is two-sided, in order to detect either positive or negative change.

To create a decision rule based on this statistic, let α denote the desired Type I error rate, i.e., α is the desired probability of rejecting the null hypothesis when it is actually true. Let $z_{1-\alpha/2}$ denote the $1-\alpha/2$ quantile of the standard normal distribution. Statistically significant change is then said to have occurred when $|Z| \geq z_{1-\alpha/2}$.

The *Z*-test is similar to the confidence interval decision rule used by Kim-Kang and Weiss (2007, 2008) in that both use the standard normal distribution to account for variability. The difference is that the confidence interval rule computes the standard error separately for $\hat{\theta}_1$ and $\hat{\theta}_2$, whereas the *Z*-test directly calculates the standard error of $\hat{\theta}_2 - \hat{\theta}_1$ under the null. In this way, the *Z*-test is specifically tailored to the hypothesis testing problem at hand, so it can be expected to enhance the statistical power.

The likelihood-ratio chi-square test. This method is adapted from a method that was described by Agresti (1996) for categorical data. Agresti defined the following likelihood-ratio statistic:

$$\Lambda = \frac{\text{maximum likelihood when parameters satisfy } H_0}{\text{maximum likelihood when parameters are unrestricted}}. \tag{11}$$

In the context of AMC, the condition "parameters satisfy H_0" is that the latent trait is constant between the two time points, i.e., $\theta_1 = \theta_2$. By definition, $\hat{\theta}_{pool}$ is the value that maximizes the likelihood under this condition. The numerator of Equation 11 is thus the likelihood of $\hat{\theta}_{pool}$ (Equation 9) for the complete data of both time points:

$$L(\hat{\theta}_{pool};\, \boldsymbol{u}_{K+L}) = \prod_{j=1}^{K+L} p_j\left(\hat{\theta}_{pool}\right)^{u_j}\left[1 - p_j\left(\hat{\theta}_{pool}\right)\right]^{1-u_j}. \tag{12}$$

In the denominator of Equation 11, the parameters are not constrained to satisfy $\theta_1 = \theta_2$. For this unrestricted case, the likelihood is maximized by computing the MLEs separately at each time point, obtaining $\hat{\theta}_1$ and $\hat{\theta}_2$. Letting $\boldsymbol{u}_K$ denote the Time 1 response vector and $\boldsymbol{u}_L$ the Time 2 response vector, the denominator of Equation 11 is then the product of the separate likelihoods. That is, the maximum unrestricted likelihood is equal to $L(\hat{\theta}_1;\, \boldsymbol{u}_K) \times L(\hat{\theta}_2;\, \boldsymbol{u}_L)$, where the two terms are calculated individually by Equation 2.

To gauge statistical significance, $-2\log(\Lambda)$ can be compared to the chi-square distribution with the appropriate degrees of freedom (Agresti, 1996). In AMC, the alternative hypothesis includes two θ values (one for each time point) whereas the null hypothesis includes only one such value; therefore, the associated significance test has one degree of freedom. The null hypothesis is rejected if and only if $-2\log(\Lambda) \geq \chi^2_{1-\alpha}$, where $\chi^2_{1-\alpha}$ is the $1-\alpha$ quantile of the chi-square distribution with one degree of freedom. Like the *Z*-test, the likelihood-ratio chi-square test directly takes the null hypothesis into account (through the numerator of Λ) and is thus tailored to the hypothesis testing paradigm.

Extension to Variable-Length Testing

Variable-length testing can be more efficient in AMC than fixed-length tests. In particular, when examinees or patients have made substantial improvement (or decline) since the previous time point, significant change might be observed after a small number of test items (Kim-Kang & Weiss, 2007, 2008). The following considers early stopping only for the Time 2 assessment; it is assumed that the Time 1 assessment has already ended, and a determination of "change" or "no change" is sought for Time 2.

The *Z*-Test

A simple stopping rule for this test would be to apply the final rejection region to all stages—that is, to cease test administration the first time that $|Z| \geq z_{1-\alpha/2}$. If this occurs at any stage of the test (including the final stage, L), the null hypothesis would be rejected; on the other hand, if $|Z| < z_{1-\alpha/2}$ for all L stages, the null hypothesis would not be rejected. Although this procedure is convenient to implement, it has an important drawback. In particular, it is likely to lower the average test length (ATL) at Time 2, but it is also likely to raise the Type I error rate. After all, by providing more opportunities for the null hypothesis to be rejected, the proportion of rejections should be expected to increase even when the null is actually true. Therefore, this method should not be used by practitioners who seek strict adherence to the α level.

To avoid such inflation of the α level, the nominal critical value, $z_{1-\alpha/2}$, should be raised in order to account for the multiple opportunities at which to reject the null. That is, for each interim stage, the null should be rejected if and only if $|Z| \geq C_1$, where $C_1 > z_{1-\alpha/2}$. At the final stage (i.e., the stage at which the maximum possible test length is reached, and the test is forced to terminate), the null is rejected if and only if $|Z| \geq C_2$, where $C_2 > z_{1-\alpha/2}$. Similar types of sequential decision rules have been used by Bartroff, Finkelman, and Lai (2008), Lai and Shih (2004), and Siegmund (1985) in other applications.

What values of C_1 and C_2 should be used? To satisfy the desired Type I error rate, these constants should be defined so that the rejection rate never exceeds α for any $\theta_1 = \theta_2$. There are many possible values that satisfy this condition; to further constrain the solution, Lai and Shih (2004) recommended that between one-third and one-half of the Type I error be "spent" on the interim stages, and the remainder be spent on the final stage. In other words, when the null hypothesis is true, the probability of stopping early to reject the null should be equal to $\varepsilon\alpha$, where $1/3 \leq \varepsilon \leq 1/2$. The probability of not rejecting the null at any interim stage, but rejecting it at the final stage, should then be equal to $(1-\varepsilon)\alpha$ in order to achieve an overall Type I error of α. The values of C_1 and C_2 that satisfy these additional conditions can be determined through simulation.

The Likelihood Ratio Statistic

Generalizing the above analysis to the likelihood-ratio chi-square test is straightforward. At each interim stage, the null hypothesis is rejected if and only if $-2\log(\Lambda) \geq D_1$; at the final stage, the null is rejected if and only if $-2\log(\Lambda) \geq D_2$. As with the *Z*-test, D_1 and D_2 are selected so that when no true change has occurred, the probability of rejecting the null at an interim stage is $\varepsilon\alpha$ and the probability of rejection at the final stage is $(1-\varepsilon)\alpha$.

It is notable that the variable-length testing procedures never stop early to make a determination of "no change;" they only stop early to reject the null hypothesis. This seeming disparity makes sense when considering that at stage l of the second time point, early stopping should be invoked only if there is substantial evidence in favor of one hypothesis over the other. Although there might be strong evidence that change *has* occurred (e.g., if $\hat{\theta}_2^l >> \hat{\theta}_1$ and both estimators are precise), it is more difficult to be confident that change *has not* occurred. After all, even if $\hat{\theta}_1$ and $\hat{\theta}_2^l$ are very close to one another, there is usually some small estimated change between the time points. Thus, an early stopping rule to determine "no change" was not examined; such a rule will be investigated in future work.

Method

Item selection methods and hypothesis tests were compared in two simulation sets. In both simulation sets, the item selection method for Time 1 was always Fisher information at $\hat{\theta}_1^k$; Time 2 item selection was conducted using both Fisher information at $\hat{\theta}_2^l$ and Kullback-Leibler information between $\hat{\theta}_2^l$ and $\hat{\theta}_{pool}^{K+l}$. The item selection methods were completely crossed with the three hypothesis tests (confidence interval overlap test, likelihood-ratio test, and *Z*-test), for a total of six procedures. The procedures are denoted as follows:

FI-CI: Fisher information item selection, confidence interval overlap test

FI-LR: Fisher information item selection, likelihood ratio test

FI-Z: Fisher information item selection, *Z*-test

KL-CI: Kullback-Leibler information item selection, confidence interval overlap test

KL-LR: Kullback-Leibler information item selection, likelihood ratio test

KL-Z: Kullback-Leibler information item selection, *Z*-test

These combinations of methods were compared based on their capability to achieve high statistical power while maintaining a Type I error rate of approximately $\alpha = 0.05$.

In each simulation set, the "no change" condition was studied at values of $\theta_1 = \theta_2$ ranging from –2 to 2, incremented by 0.5. Power was studied at three levels of improvement (true change of θ = 0.5, 1.0, and 1.5), with the latent traits again ranging from –2 to 2. All combinations of θ_1 and θ_2 considered are indicated in Table 1, where "N" = no change, "L" = low change (0.5), "M" = medium change (1.0), and "H" = high change (1.5). 1,000 replications were performed for every combination of θ_1 and θ_2 and both item selection methods; the three hypothesis tests were then applied to each set of 1,000 simulees.

Table 1. Combinations of θ_1 and θ_2 Studied in Each Simulation Set

Time 1	Time 2 (θ_2)								
(θ_1)	–2.0	–1.5	–1.0	–0.5	0.0	0.5	1.0	1.5	2.0
–2.0	N	L	M	H	--	--	--	--	--
–1.5	--	N	L	M	H	--	--	--	--
–1.0	--	--	N	L	M	H	--	--	--
–0.5	--	--	--	N	L	M	H	--	--
0.0	--	--	--	--	N	L	M	H	--
0.5	--	--	--	--	--	N	L	M	H
1.0	--	--	--	--	--	--	N	L	M
1.5	--	--	--	--	--	--	--	N	L
2.0	--	--	--	--	--	--	--	--	N

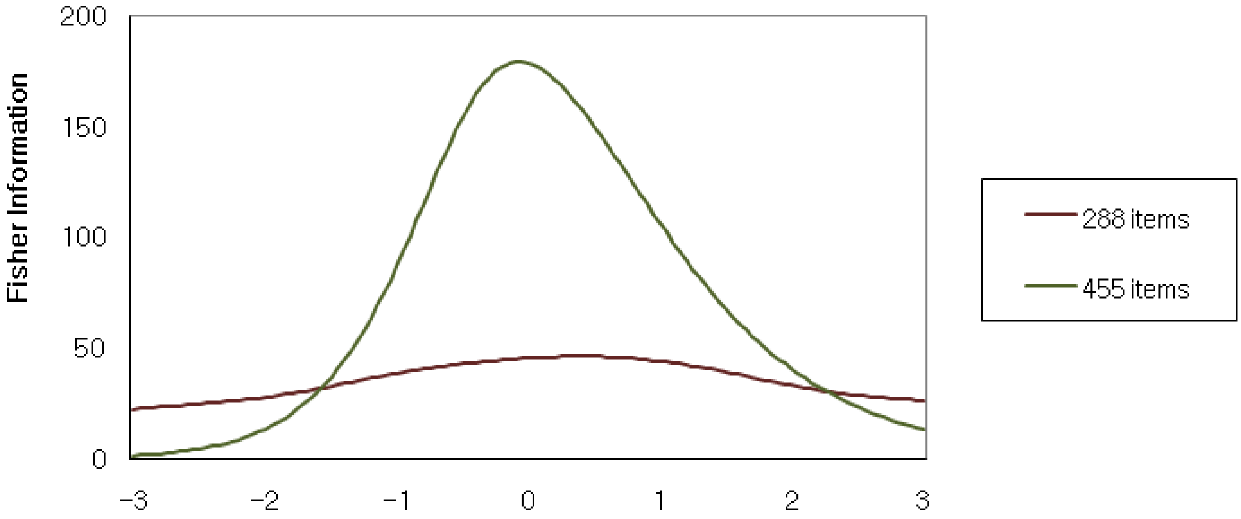
N = no change, L = low change (0.5), M = medium change (1.0), H = high change (1.5).

As in Kim-Kang and Weiss (2007, 2008), item selection methods were compared in their unconstrained form. Therefore, exposure control (e.g., Chang, Qian, & Ying, 2001; Stocking & Lewis, 1998; Sympson & Hetter, 1985) and content balance (e.g., Kingsbury& Zara, 1989; van der Linden, 2000) were not applied, and item eligibility rules for the two time points were defined by option 3 above. These liberal conditions are realistic in many psychological assessments, as well as some low-stakes educational assessments.

The item bank for Simulation Set 1 was the relatively ideal CAT item bank used in the "medium discrimination" condition of Kim-Kang and Weiss (2007, 2008). It consisted of 288 simulated items. The a_j parameters were simulated from the normal distribution with a mean of 1 and a standard deviation of 0.15. The b_j parameters were simulated so that a specified number would fall into each of 18 intervals ranging from [–4.5, –4.0] to [4.0, 4.5]. In particular, 24 items were located in each of the six middle intervals ([–1.5, –1.0] to [1.0, 1.5]) and 12 items were located in each of the outer intervals ([–4.5, –4.0] to [–2.0, –1.5] and [1.5, 2.0] to [4.0, 4.5]). Within each interval, the b_j parameter was simulated from the uniform distribution. This procedure created an adequate number of items in the middle of the θ distribution; it also ensured coverage beyond the range of the true θ values under study (–2 to 2). Finally, c_j was set at 0.20 for all 288 items (Kingsbury & Weiss, 1983; Lord & Novick, 1968; Urry, 1997; Yen, 1986). Simulees were administered 50 items at each time point in Simulation Set 1.

Simulation Set 2 used a more realistic bank of 455 items that were candidates for a statewide English Language Arts examination of Grade 10 students. The mean and standard deviation of a_j values across the bank were 1.02 and 0.34, respectively; those of b_j values were 0.02 and 0.78; those of c_j values were 0.22 and 0.06. Figure 1 shows the respective information functions of the two item banks. To evaluate the different AMC procedures under a shorter test length, simulees were administered only 30 items at each time point in Simulation Set 2.

Figure 1. Fisher Information Function for the Item Banks



At each time point and for each simulation set, the MLE was bounded in the range [–4, 4]. For instance, if all items were answered correctly up to a given stage, the MLE was set to 4 rather than taking on an infinite value. This rule facilitated item selection at the early stages of the test.

Results

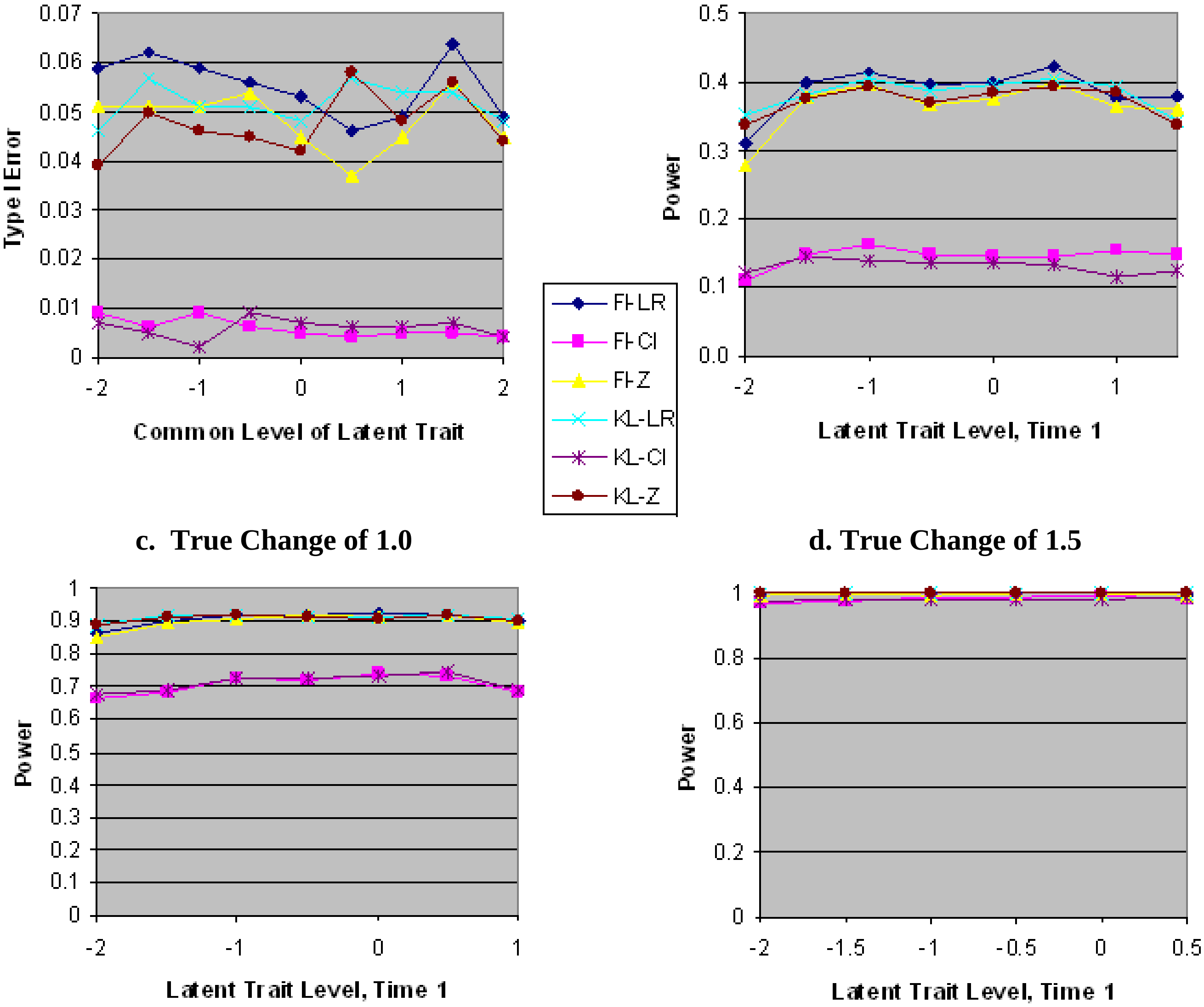
Fixed-Length Tests

Simulation set 1. Results of Simulation Set 1 are presented in Figure 2. Figure 2a displays the Type I error rate of all six procedures, plotted against the true value of $\theta_1 = \theta_2$, under the "no change" condition. Figures 2b–2d plot the statistical power of these procedures against θ_1 under true change values of θ = 0.5, 1.0, and 1.5, respectively.

The most salient feature of Figure 2 involves the two procedures using the confidence interval overlap hypothesis test, namely FI-CI and KL-CI. Although the desired Type I error rate was $\alpha = 0.05,$ the observed Type I error rate of these procedures never reached 0.01 for any level of the latent trait (Figure 2a). Having a Type I error rate far below the intended value is not problematic in itself, but it indicates the conservative nature of the CI overlap test. Such a conservative approach resulted in low power for the FI-CI and KL-CI methods at true change values of 0.5 and 1.0 (Figures 2b and 2c). On the other hand, the four procedures using the LR test or *Z*-test generally exhibited Type I error rates between 0.04 and 0.06, close to the desired value of $\alpha = 0.05.$ The power functions of these procedures greatly surpassed those of FI-CI and KL-CI for true change values 0.5 and 1.0, with gains ranging from 0.16 to 0.29. At the true change value of 1.5, the power of all methods was approximately 1 (Figure 2d).

Discounting FI-CI and KL-CI due to inadequate power, the remaining four procedures (FI-LR, FI-Z, KL-LR, and KL-Z) were compared. Although their performances were relatively similar, some differences are notable. FI-LR and KL-LR tended to exhibit the highest power, but also the highest Type I error rates. In particular, FI-LR had the highest Type I error rate in seven of nine "no change" conditions; KL-LR had the highest or second-highest Type I error rate in six of nine such conditions. KL-Z tended to maintain the desired Type I error rate, only exceeding 0.05 in two of nine “no change” conditions (as opposed to six, six, and five conditions over 0.05 for FI-LR, KL-LR, and FI-Z, respectively). In fact, among these four procedures, KL-Z had the lowest Type I error rate in six of nine “no change” conditions, yet in most “positive change” conditions displayed more power than FI-Z and power within 1% of FI-LR and KL-LR. Thus, although no procedure uniformly outperformed its competitors, KL-Z tended to exhibit the best overall classification properties, albeit by a slight margin.

Figure 2. Comparison of Type I Error Rates and Power Functions, Simulation Set 1



a. No Change

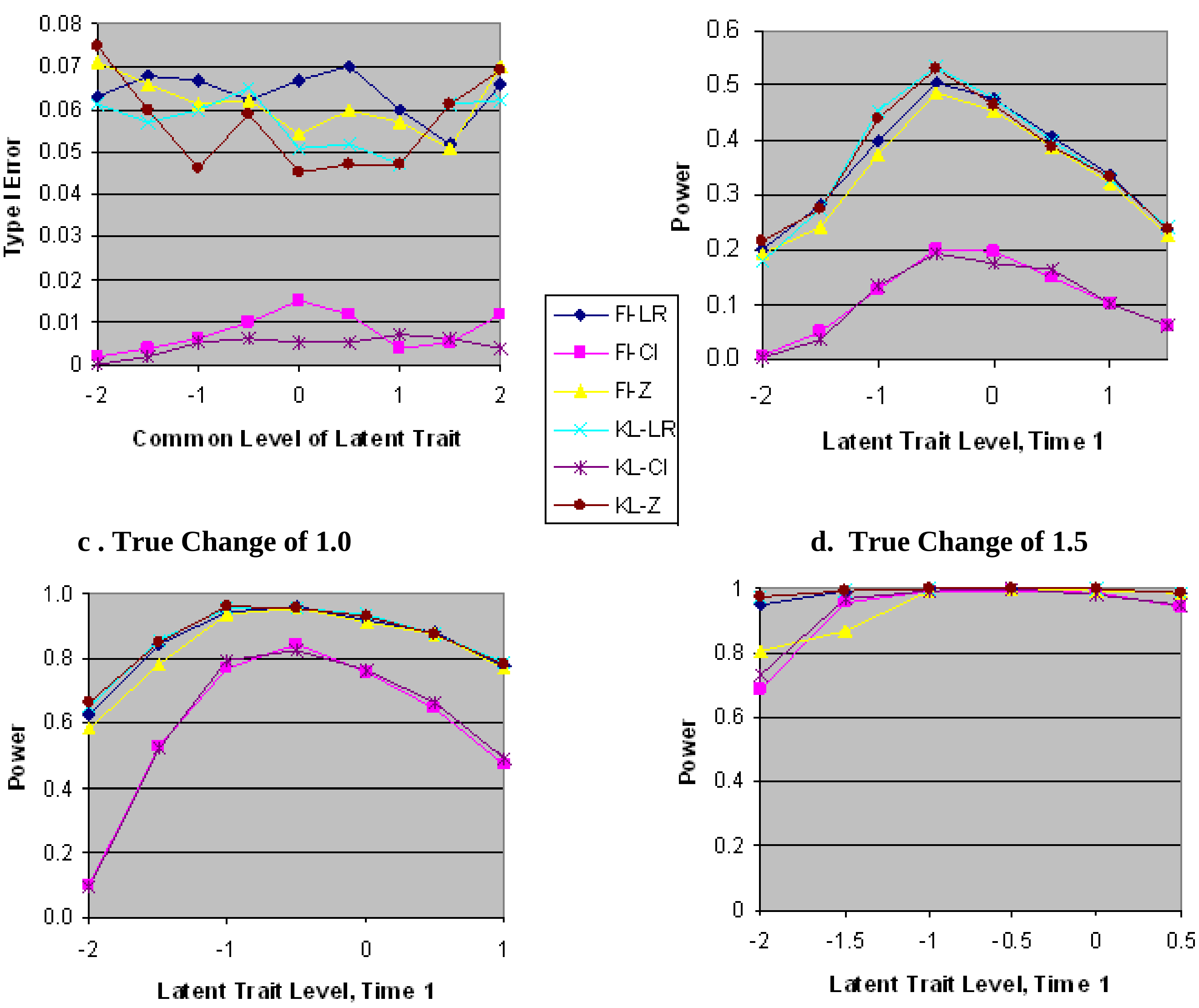
b. True Change of 0.5

c. True Change of 1.0

d. True Change of 1.5

Simulation set 2. Figure 3 presents the results for Simulation Set 2. As in Simulation Set 1, FI-CI and KL-CI had Type I error rates far below 0.05 (Figure 3a). This again resulted in low power for these methods, particularly at a positive change of 0.5 (Figures 3b–3d). The other four procedures all displayed far greater power than FI-CI and KL-CI; they also exhibited slight inflation of the Type I error rate, though this rate exceeded 0.07 on just two occasions. KL-Z had a relatively high Type I error rate at extreme values such as $\theta_1 = \theta_2 = \pm 2,$ but it was the only method in addition to FI-CI and KL-CI to maintain the desired Type I error rate in more than one condition, doing so four times. Discounting the CI methods for lack of power, KL-Z had the lowest Type I error rate in five of nine "no change" conditions and also exhibited competitive power. Thus, the findings were similar to Simulation Set 1: no method uniformly outperformed its competitors, but KL-Z had the best balance of Type I error and power.

Figure 3. Comparison of Type I Error Rates and Power Functions, Simulation Set 2



a. No Change

b. True Change of 0.5

c . True Change of 1.0

d. True Change of 1.5

A general comparison of Simulation Sets 1 and 2 indicates the effect of the item bank on the classification properties of AMC. Even with a test of 20 fewer items, Simulation Set 2 displayed enhanced power in the low-to-middle portion of the θ continuum due to the greater information in that region. However, the item bank of Simulation Set 1 had more items located at the extremes of the continuum (Figure 1) and therefore exhibited better power there. Thus, these simulations confirmed the anticipated result that the characteristics of the item bank are an important factor in the quality of AMC, as they are in any CAT application.

Finally, at a positive change of 0.5, the power never exceeded 0.422 in Simulation Set 1 or 0.534 in Simulation Set 2. Kim-Kang and Weiss (2007, 2008) illustrated the effectiveness of AMC relative to non-adaptive methods for measuring change, so these somewhat low power values do not represent a lack of efficiency in AMC. Rather, they reflect the inherent difficulty of comparing two performances that are both measured with error, whether done adaptively or using a linear form.

Variable-Length Tests

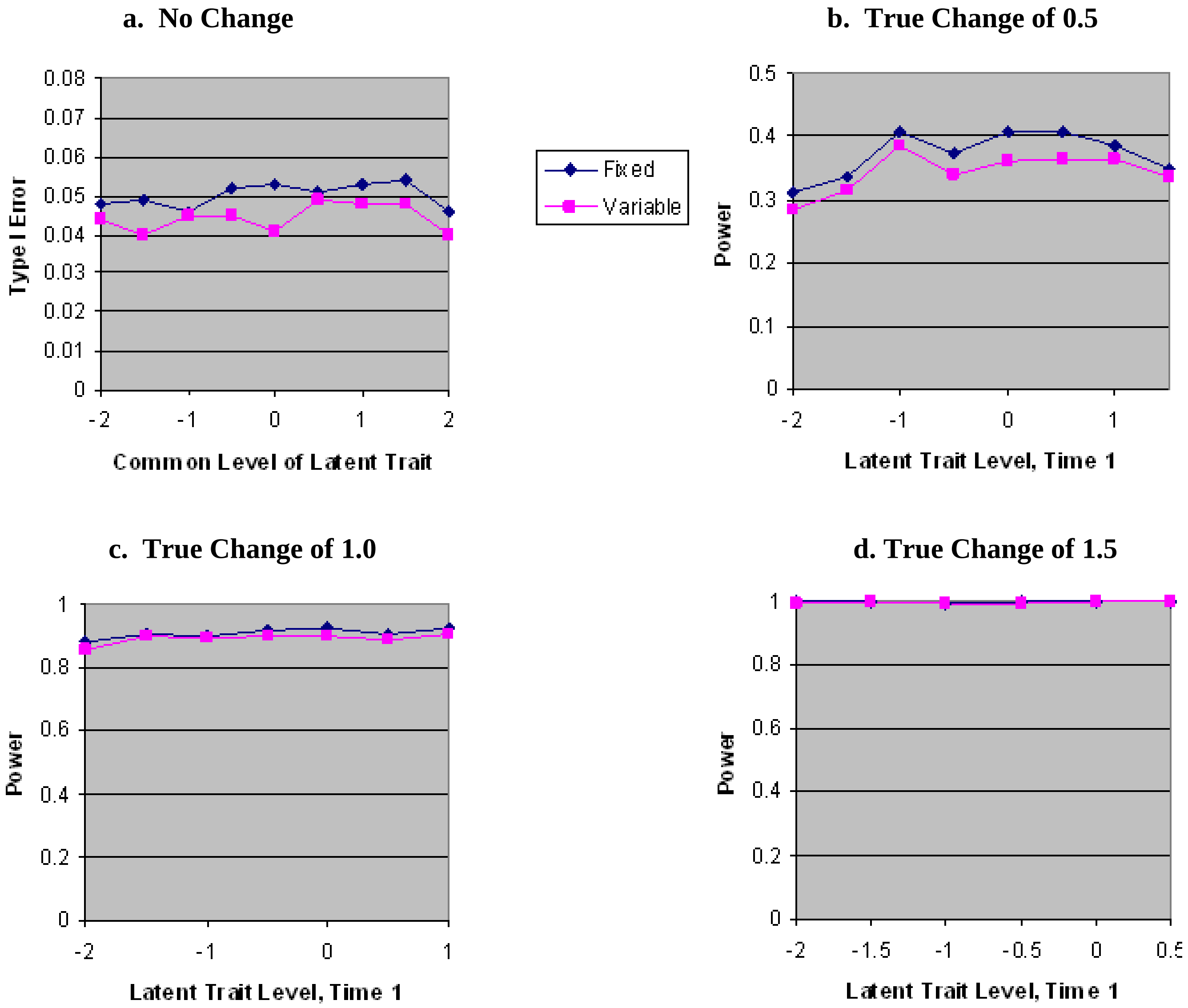
To investigate the degree to which efficiency can be improved through variable-length testing, a simulation was run comparing a fixed-length procedure to its variable-length extension. For brevity, only one simulation set and one AMC procedure were examined. Simulation Set 1 was chosen because its fixed-length design called for 50 items rather than 30, allowing for greater potential savings in test length. KL-Z was chosen because this procedure had performed well in the fixed-length study.

The fixed-length test (FLT) administered 50 items at each time point, as was done for the other simulations. The variable-length test (VLT) also administered a uniform 50 items at Time 1, but allowed for different stopping times to occur at Time 2. For face validity, as well as to avoid early stopping based on an unstable θ estimate, a minimum test length of 20 was set for Time 2. The maximum test length was set at 50 so that the VLT was never longer than the FLT. Preliminary simulations indicated that $C_1 = 2.67$ and $C_2 = 2.07$ were appropriate critical values for the VLT.

Type I error rates and power functions are presented in Figure 4. As C_1 and C_2 had been specifically selected to achieve a Type I error rate of $\alpha = 0.05$ or lower, the VLT's observed α level never exceeded this desired threshold (Figure 4a). On the other hand, the critical value for the fixed-length test, 1.96, was obtained from the normal approximation; hence, although the FLT's observed α level was always close to 0.05, it sometimes exceeded this value. Because the FLT's observed α level was higher than that of the VLT for every "no change" condition studied, its power would be expected to be higher. This expectation was borne out in the results: the FLT had greater power, ranging from 1.3% to 4.4% at a positive change of 0.5, from 0.8% to 2.5% at a positive change of 1.0, and 0.0% to 0.1% at a positive change of 1.5 (Figures 4b–4d). Note, however, that the VLT's classification properties reported above (better Type I error and worse power than the FLT) are partially an artifact of the chosen C_1 and C_2; the classification properties of the VLT and FLT could be made more similar by either lowering these values or raising the FLT's critical value above 1.96.

Figure 5 shows the mean and standard deviation of the number of fewer items administered by the VLT than by the FLT (note the changes in scale across the four panels). In the "no change" condition, mean savings ranged from 0.4 to 0.6; corresponding ranges were 3.7 to 6.5, 19.0 to 21.4, and 28.6 to 29.0 items for positive change values of 0.5, 1.0, and 1.5, respectively. Thus, the VLT exhibited moderate reductions in the average test length with a change of 0.5, and substantial reductions at 1.0 and 1.5. As indicated by the standard deviations, there was substantial variability in item savings when the assumed change value was 0.5 or 1.0.

Figure 4. Comparison of Type I Error Rates and Power Functions for Fixed-Length and Variable-Length Tests Using KL-Z



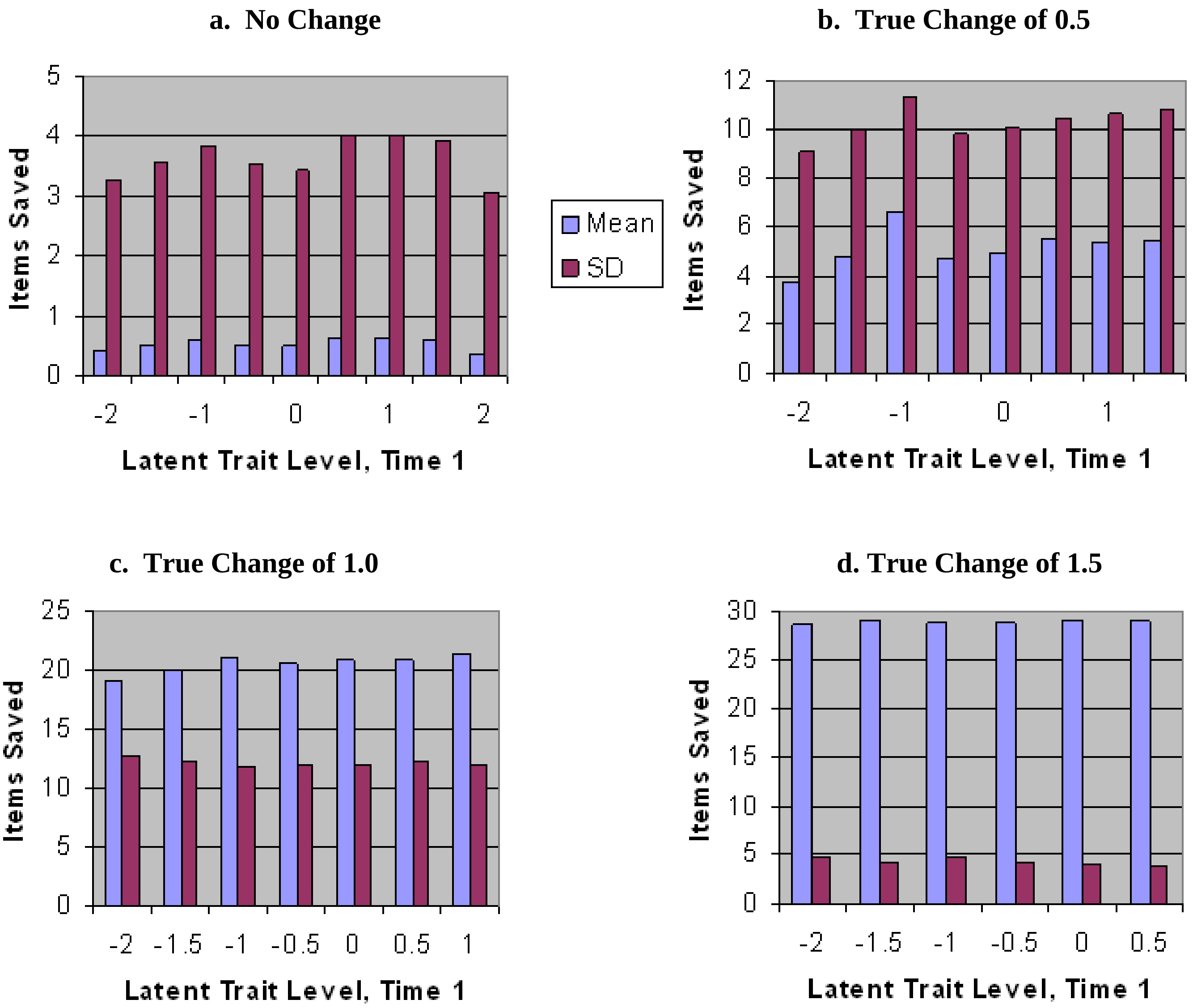
a. No Change

b. True Change of 0.5

c. True Change of 1.0

d. True Change of 1.5

Figure 5. Number of Items Saved in Variable-Length Testing Using KL-Z



a. No Change

b. True Change of 0.5

c. True Change of 1.0

d. True Change of 1.5

Discussion and Conclusions

The goal of this study was to further investigate the capability of the AMC method to detect significant change by (1) introducing a new item selection criterion, (2) implementing two hypothesis testing methods, and (3) exploring the use of variable-length testing in this context. The simulation results using fixed-length tests supported the greater efficiency of the new procedures over existing methods. Both the *Z*-test and LR test exhibited much higher power than the method based on overlapping confidence intervals. The difference in performance among item selection methods was more subtle, but using a Kullback-Leibler information criterion at Time 2 resulted in a slight advantage over the use of Fisher information. Overall, the combination of Kullback-Leibler item selection and the *Z*-test displayed a good balance of Type I error and power in both simulation sets. This combination was re-run with variable-length testing at Time 2, which was found to yield substantial savings in average test length as compared to fixed-length AMC in the detection of significant change.

This research used monte-carlo simulation methods. Any monte-carlo simulation can be questioned with respect to whether its assumptions are realistic. In this research, the primary assumptions were the nature of the item banks used and the magnitudes of change examined.

Two item banks were used. Bank 1 was an "ideal" CAT item bank with item difficulties well-distributed throughout the θ range and moderately discriminating items. Bank 2 came from an operational item bank. Results showed that the power of the ideal bank was more constant across the θ distribution than that of the operational bank. The operational bank displayed less power at the extremes of the continuum but more power in the low-to-middle portion of the θ range. Thus, these simulations confirmed the anticipated result that the characteristics of the item bank are an important factor in the quality of AMC, as they are in any CAT application.

Three magnitudes of individual change were simulated in this study: θ = 0.5, 1.0, and 1.5. Justification for choice of these change values is difficult because there is little individual change data available in the IRT metric. However, change data in a classical number- or percent-correct metric can approximate the IRT standard score metric. Table 2 shows alternate forms retest data for students in three grades from schools in the U.S. Because the focus of this research was on *individual* change (rather than group change), change was computed as the simple difference score for each student of their Time 2 (Form B) percent correct score test minus their Time 1 (Form A) percent correct score. The mean change score was then divided by the Time 1 standard deviation (SD) to express the average change on a *z*-score scale (i.e., change was expressed in Time 1 SD units, as was done in the simulation). In addition, the maximum observed percent change score was divided by the Time 1 score SD, to express maximum observed change in Time 1 SD units.

Table 2. Mean and SD of Time 1 (Form A) Achievement Test Percent Correct Scores, Mean and SD of Individual Change Scores (Form B minus Form A), and Mean and Maximum of Individual Change Scores Expressed in Time 1 SD Units (Tests Ranged From 40 to 50 Items With Retest Over an Approximately 4.5-Month Interval)

			Time 1		Change		Change in Time 1 SD Units	
Subject	Grade	*N*	Mean	SD	Mean	SD	Mean	Maximum
Math	3	268	42.47	13.21	18.57	13.30	1.406	4.39
Science	3	192	40.96	16.01	21.44	15.57	1.339	4.16
Math	4	177	42.99	15.27	12.32	11.54	0.807	3.27
Reading	4	176	57.32	19.14	-1.99	11.72	-0.104	1.67
Math	5	182	49.18	17.40	10.27	12.74	0.590	3.79
Reading	5	179	55.94	18.32	0.24	5.74	0.001	5.00
Science	5	179	50.65	16.90	5.32	13.22	0.315	2.77
Social Studies	5	178	46.58	16.25	12.56	10.66	0.773	2.83

Table 2 shows a range of mean individual change from 1.406 SD units to 0.001 SD units across the eight data sets. The five largest means of 1.406, 1.339, 0.807, 0.773, and 0.590 support the use of all three simulated change levels used in this study. Note that approximately half the observed individual change values in these data were higher than the deviated means shown in

Table 2. The maximum individual deviated change ranged from a low of 1.67 to a high of 5.00, with all but the first above 2.5 SD units. For two groups—Reading, grades 4 and 5—mean deviated change was about zero, indicating no change *on average* in these data. However, there were obviously individual students with substantial amounts of change. Because the AMC procedure is concerned with identifying *individual* change, the observed maximum levels of deviated change shown in Table 2 readily support the levels of individual change used in these simulations.

The procedures introduced herein were designed to enhance the advantages of AMC over non-adaptive methods to detect and measure change. Such advantages are detailed in Kim-Kang and Weiss (2007, 2008). Many more extensions of AMC research are needed, including:

1. Comparison of item selection procedures at the first time point. For instance, Fisher information might be compared with the original Chang and Ying (1996) version of Kullback-Leibler information.
2. Variable-length testing at the first time point, such as stopping early when the standard error of measurement falls below a certain threshold.
3. A variable-length method that can be used at Time 2 to terminate testing when no significant change has occurred.
4. Further examination of the “minimum test length” requirement used here. Results showed that a large change can frequently be detected quickly. Other minimum test length values should be investigated.
5. Extension to more than two time points. For example, suppose that measurements are to be taken at three time points. An effective procedure would exhibit high power for detecting change between Times 1 and 2 simultaneously allowing for comparison at Time 3. Hence, even if there is adequate evidence that $\theta_1 \neq \theta_2$, the procedure might continue testing at Time 2 until a low standard error of measurement has been achieved to allow for the detection of further change between Time 2 and Time 3.
6. Application of AMC to measurement with polytomous models (e.g., Muraki, 1992; Samejima, 1969). Such models are often used in psychological assessments.
7. Item selection incorporating exposure control, content balance, and constraints preventing an examinee from being administered the same item at more than one time point. Such considerations are typically more important in educational assessment than psychological assessment.

The above list illustrates that AMC is a fertile area of research. It is also an inherently challenging area due to the fact that measurement at every time point is made with error. The statistical power of AMC (and any other method for detecting change) is likely adversely affected by these multiple sources of error. Nevertheless, the tracking of progress in both students and patients is an important application of psychometrics, and one that merits further study in the future. The AMC paradigm provides a solution that is not available using conventional fixed-form tests.

References

Agresti, A. (1996). *An introduction to categorical data analysis.* New York: John Wiley & Sons, Inc.

Baker, F. B., & Kim, S-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Marcel Dekker.

Bartroff, J., Finkelman, M., & Lai, T. L. (2008). Modern sequential analysis and its applications to computerized adaptive testing. *Psychometrika, 73,* 473-486.

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In C. Harris (Ed.), *Problems in measuring change* (pp. 3-20). Madison, WI: University of Wisconsin Press.

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

Burr, J. A., & Nesselroade, J. R. (1990). Change measurement. In A. von Eye (Ed.), *Statistical methods in longitudinal research* (vol. 1) (pp. 3-34). Boston, MA: Academic Press.

Chang, H. H., & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika, 58,* 37-52.

Chang, H-H., Qian, J., & Ying, Z. (2001). *a*-stratified multistage CAT with *b*-blocking. *Applied Psychological Measurement, 25,* 333-341.

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

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory.* New York: John Wiley & Sons, Inc.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change"—or should we? *Psychological Bulletin, 74,* 68-80.

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23,* 249-261.

Embretson, S. E. (1995). A measurement model for linking individual learning to processes and knowledge: Application to mathematical reasoning. *Journal of Educational Measurement, 32,* 277-294.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park: SAGE Publications.

Kim-Kang, G., & Weiss, D. J. (2007). Comparison of computerized adaptive testing and classical methods for measuring individual change. In D. J. Weiss (Ed.), *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.* Available from www.psych.umn.edu/psylabs/CATCentral/

Kim-Kang, G., & Weiss, D. J. (2008). Adaptive measurement of individual change. *Zeitschrift fur Psychologie / Journal of Psychology, 216,* 49-58.

Kingsbury, G. G., & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing: Computerized adaptive testing and latent trait test theory* (pp. 257-283). New York: Academic Press.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2,* 359-375.

Lai, T. L., & Shih, M. C. (2004). Power, sample size and adaptation considerations in the design of group sequential trials. *Biometrika, 91*, 507-528.

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Menlo Park, CA: Addison-Wesley.

Manning,W. H., & DuBois, P. H. (1962). Correlation methods in research on human learning. *Perceptual and Motor Skills*, *15*, 287-321.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159–176.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, Monograph Supplement No. 17.

Siegmund, D. (1985). *Sequential analysis: Tests and confidence intervals.* New York: Springer-Verlag.

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.

Sympson, J. B., & Hetter, R. D. (1985). Controlling item exposure rates in computerized adaptive testing. In *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 937-977). San Diego, CA: Navy Personnel Research and Development Center.

Urry, V. W. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement, 14*, 181-196.

van der Linden, W. J. (2000). Constrained adaptive testing with shadow tests. In W.J. van der Linden & C. A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 27-52). Boston, MA: Kluwer.

van der Linden, W. J., & Glas, C. A. W. (Eds.). (2000). *Computerized adaptive testing: Theory and practice.* Boston, MA: Kluwer.

Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A Primer (2nd Edition)*. Mahwah, NJ: Erlbaum.

Weiss, D. J. (1983). Introduction. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 1-9). New York: Academic Press.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21,* 361-375.

Willett, J. B. (1997). Measuring change: What individual growth modeling buys you. In E. Amsel and K.A. Renninger (Ed.), *Change and development: Issues of theory, method, and application* (pp. 213-243). Mahwah, NJ: Erlbaum.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23,* 299-325.