Item Selection and Hypothesis Testing for the Adaptive Measurement of Change

Matthew Finkelman, Harvard School of Public Health

David J. Weiss, University of Minnesota

Gyenam Kim-Kang, Korea Nazarene University

In press, Applied Psychological Measurement, 2010

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.

Keywords: change, individual change, measuring change, computerized adaptive testing, likelihood ratio, Kullback-Leibler information.

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 (or the same) 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 determination of change is again made on the basis of responses to test items, resulting in test scores with varying amounts of error, 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, there have been a number of attempts to address the problem.

One traditional approach to measuring individual change is to compute the simple difference between scores obtained on two (or more) occasions (e.g., Time 1 and Time 2), such as in a pretest-posttest paradigm (e.g., Burr & Nesselroade, 1990; McDonald, 1999). Previous research regarding this simple difference score for the measurement of individual change has

demonstrated that it has major problems, including low reliability (Embretson, 1995; Hummel-Rossi & Weinberg, 1975; Lord, 1963; Willett, 1994, 1997), negative correlation between change scores and initial status (Cronbach & Furby, 1970; Embretson, 1995; Willett, 1994, 1997), regression toward the mean (Cronbach & Furby, 1970; Hummel-Rossi & Weinberg, 1975), and dependence on potentially different scales (Embretson, 1995; Hummel-Rossi & Weinberg, 1975).

Other procedures for estimating change have been suggested and examined (Lord, 1963; McNemar, 1958; Tucker, Damarin, & Messick, 1966). The residual change score (RCS), proposed by Manning and DuBois (1962), is one of the most frequently advocated alternatives to the simple difference score (Willett, 1994, 1997). The RCS reflects the difference between an observed score at Time 2 and a predicted score based on linear regression. Manning and DuBois (1962) showed analytically that the RCS typically is more reliable than the simple difference score. However, the RCS is dependent on group level information and is not the actual amount of change, but indicates how much different the observed score at Time 2 is from the value that is predicted. Consequently, the RCS is appropriate for studying the correlates of change, but not for evaluation of individual change (Kim-Kang & Weiss, 2007, 2008).

Some of the problems in earlier attempts to measure change were due to the use of classical test theory (CTT) for developing measuring instruments to be used in measuring individual change. In CTT, tests are usually developed to maximize internal consistency reliability, by selecting items with difficulties around $p = .50$. This results in tests that measure well around the difficulty level of the test, but measure poorly elsewhere; such tests are suboptimal for measuring change. In addition, CTT item statistics – difficulty and discrimination – as well as its test scores are dependent on the samples of persons and items on

which they are computed. These characteristics also mitigate against the successful measurement of individual change.

In recent years, item response theory (IRT; Embretson & Reise, 2000; Lord, 1980) has begun to replace CTT for the development of measuring instruments. IRT has several advantages over CTT for measuring individual change. IRT can place different examinees (or the same examinee at different time points) on the same scale when different items are used. Similarly, item parameters estimated on a sample of examinees at one trait level can be transformed to or linked with those at another trait level, thereby allowing the development of item banks that can cover a wide trait range. Using IRT procedures, tests can be assembled at different levels along this trait continuum, thus providing a scale along which change can be measured.

A number of researchers have addressed the issue of measuring change using IRT models, including the linear logistic latent trait models (Fischer, 1976), an IRT model for growth curves (Bock, 1976), the multidimensional Rasch model for repeated testings (Andersen, 1985), and a multidimensional Rasch model for learning and change (Embretson, 1991a, 1991b). The model proposed by Embretson is the only IRT model that provides change parameters for measuring individual change, but it is restricted to a one-parameter logistic multidimensional IRT model that requires the unrealistic assumption of equal discriminations across items. Fischer's linear logistic IRT model estimates group change, Bock's model requires group level information, and Andersen's approach is designed to assess the relationship between the latent trait at two time points and/or changes in the latent trait across time rather than to estimate the extent of individual change.

A recent book on analyzing change (Collins & Sayer, 2001) addresses contemporary developments in the field, with a focus on intraindividual variability. Although a variety of potentially useful approaches for analyzing change data are described, including time series, dynamical, and multilevel models, very little attention is focused on the quality of the measurements used in the analytic procedures – there is little or no consideration of whether the observed measurements actually reflect true change or whether significant change can be identified for a given individual. Thus, a different approach is needed that is designed to accurately measure and identify significant individual change when it exists.

Kim-Kang and Weiss (2007, 2008) combined the modern technologies of IRT and computerized adaptive testing (CAT) in a procedure, originally proposed by Weiss and Kingsbury (1984), that they called adaptive measurement of change (AMC). Using Monte Carlo simulation, they then compared four procedures for measuring individual change: the simple difference score (Burr & Nesselroade, 1990; McDonald, 1999), the RCS, a difference score based on IRT, and AMC. Their AMC procedure (described in detail below) better captured actual levels of change than did the other methods and was able to efficiently detect true change. 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 goal is to *estimate* change, but might be improved upon when the goal is a powerful *hypothesis test* of change. Additionally, their approach to detecting change had low power 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 for AMC. These item selection and hypothesis testing methods were compared to the original AMC procedure in simulation.

The AMC Procedure

CAT (van der Linden & Glas, 2000; Wainer, 2000; Weiss, 1983; Weiss & Kingsbury, 1984), which uses IRT's capability of placing different examinees (or the same examinee at different time points) on the same scale even when different items are used across measurements, tailors an instrument 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 Time 1 and the other at Time 2. Let θ_1 denote the value of an 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 1 to Time 2. 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. Most CAT item selection methods require that an *interim* θ *estimate* be made following each item response. After k items have been presented, eliciting a response vector $\mathbf{u}_k = (u_1, ..., u_k)$, a common choice is the maximum likelihood estimate (MLE) of θ; this will be denoted $\hat{\theta}_k$. CAT usually then selects the item that maximizes the Fisher information at $\hat{\theta}_k$, among all items that have not yet been administered to the examinee (e.g., Lord, 1980, p. 153). When the MLE does not exist, an item is typically selected to maximize the Fisher information at an arbitrary 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 estimated θ separately for each measurement. That is, at Time 1, item $k+1$ was selected to

maximize $I_j(\hat{\theta}_1^k)$, where $\hat{\theta}_1^k$ denotes the Time 1 MLE after k items and I_j denotes the information function. 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 $\hat{\theta}_2$.

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 $\hat{\theta}_1 \pm z_{1-\alpha/2}\, SE\left(\hat{\theta}_1\right)$, where $z_{1-\alpha/2}$ is the appropriate quantile of the standard normal distribution and $SE\left(\hat{\theta}_1\right)$ is the observed standard error of measurement evaluated at $\hat{\theta}_1$, computed by

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

where $\left[I\left(\hat{\theta}_1\right)\right]$ is observed information obtained from the second derivative of the likelihood function. The confidence interval for θ_2 is analogous. Statistically significant change was identified if these intervals did not overlap.

New Methods for AMC

Item Selection At Time 2

The item selection method of Kim-Kang and Weiss (2007, 2008) is a reasonable approach for AMC when the goal is estimation. However, in the current application of AMC, the power to detect true change might be enhanced by selecting Time 2 items to differentiate between θ_1 and θ_2. 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 uses Kullback-Leibler information (KLI; Cover & Thomas, 1991), which 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 KLI of item j for distinguishing these values is equal to

$$K_j(\theta', \theta'') = E_{\theta'}\left[\ln \frac{L\left(\theta'; u_j\right)}{L\left(\theta''; u_j\right)}\right], \tag{2}$$

where $E_{\theta'}\left[X\right]$ denotes the expectation of X under θ'. A large value of $K_j\left(\theta', \theta''\right)$ 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 $\boldsymbol{u}_{K+l}$ 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" or $\hat{\theta}_{pool}^{K+l}$) is the value of θ maximizing

$$L(\theta; \boldsymbol{u}_{K+l}) = \prod_{j=1}^{K+l} p_j(\theta)^{u_j} \left[1 - p_j(\theta)\right]^{1-u_j}, \tag{3}$$

where $p_j(\theta)$ is the probability of selecting the correct/keyed response under an appropriate IRT model.

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 $K_j(\hat{\theta}_2^l, \hat{\theta}_{pool}^{K+l})$. [See Eggen (1999) for a related application of KLI in the context of computerized classification testing.] 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 KLI 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.

Hypothesis Testing

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

$$|Z| = \frac{|\hat{\theta}_2 - \hat{\theta}_1|}{\sqrt{\frac{1}{\sum_{j=1}^{K} I_j\left(\hat{\theta}_{pool}\right)} + \frac{1}{\sum_{j=1}^{L} I_j\left(\hat{\theta}_{pool}\right)}}}, \quad (4)$$

where $\hat{\theta}_{pool}$ is the pooled estimate after both tests have been completed, $\sum_{j=1}^{K} I_j\left(\hat{\theta}_{pool}\right)$ 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 information can be used to approximate the variance of the MLE (Chang & Stout, 1993); here, information is evaluated at $\hat{\theta}_{pool}$ because this value is the most plausible θ level under H_0. By the additivity of variances under local independence, the denominator in Equation 4 can be considered the standard error of $\hat{\theta}_2 - \hat{\theta}_1$ under H_0. 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. 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 H_0.

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:

$$\Lambda = \frac{\text{maximum likelihood when parameters satisfy } H_0}{\text{maximum likelihood when parameters are unrestricted}}. \quad (5)$$

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 5 is thus the likelihood of $\hat{\theta}_{pool}$ for the complete data of both time points.

In the denominator of Equation 5, 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 5 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 separately.

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. This procedure would be likely to lower the average test length (ATL) at Time 2, but also to raise the Type I error rate: 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 if strict adherence to the α level is desired.

To avoid such inflation of the α level, the nominal critical value, $z_{1-\alpha/2}$, should be raised. 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.

To satisfy the desired Type I error rate, C_1 and C_2 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; 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 $1/3 \le \varepsilon \le 1/2$. 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. For the likelihood-ratio chi-square test, the null hypothesis is rejected during interim stages if $-2\log(\Lambda) \ge D_1$; at the final stage, the null is rejected if $-2\log(\Lambda) \ge D_2$. 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 $(1-\varepsilon)\alpha$.

These variable-length testing procedures never stop early to make a determination of "no change;" they stop early only to reject the null hypothesis. This seeming disparity makes sense when considering that at stage l of Time 2, 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: even if $\hat{\theta}_1$ and $\hat{\theta}_2^l$ are very close to each

other, there is usually some small estimated change between the time points. Thus, an early stopping rule to determine "no change" was not examined.

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 (FI) at $\hat{\theta}_1^k$; Time 2 item selection was conducted using both FI at $\hat{\theta}_2^l$ and KLI 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 (CI), likelihood-ratio test (LR), and Z-test (Z)—for a total of six procedures: FI-CI, FI-LR, FI-Z, KL-CI, KL-LR, and KL-Z. These combinations of methods were compared based on their statistical power and Type I error rates.

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.

Insert Table 1 about here

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.

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 (discrimination) parameters of the 3-parameter logistic model (Lord, 1980, p. 12) were simulated from the normal distribution with a mean of 1 and a standard deviation (SD) of 0.15. The b_j (location) 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 (the pseudo-guessing parameter) was set at 0.20 for all 288 items (Lord & Novick, 1968; 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 from a statewide English Language Arts examination of Grade 10 students. The four-option multiple-choice items tested the students' level of reading comprehension. The mean and SD of a_j values across the bank were 1.02 and 0.34, respectively; 0.02 and 0.78 for b_j; and 0.22 and 0.06 for c_j. Figure 1 shows the 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. At each time point and for each simulation set, the MLE was bounded in the range [–4, 4].

Insert Figure 1 about here

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.

Insert Figure 2 about here

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, which 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 considerably 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).

Although the performance of the new procedures (FI-LR, FI-Z, KL-LR, and KL-Z) was relatively similar, some differences were 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 for seven of nine θ levels; KL-LR had the highest or second highest Type I error rate at six of nine θ levels. KL-Z tended to maintain the desired Type I error rate, exceeding 0.05 for only two of nine θ levels in the "no change" condition (as opposed to five or six for the other methods). Among these four procedures, KL-Z had the lowest Type I error rate for six of nine θ levels, 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 the others, KL-Z tended to exhibit the best overall classification properties, albeit by a slight margin.

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 with change of 0.5 (Figure 3b). 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 at more than one θ level, doing so four times. Discounting the CI methods for lack of power, KL-Z had the lowest Type I error rate at five of nine θ levels 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.

Insert Figure 3 about here

A 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) with consequent better power there.

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 test.

Variable-Length Tests

For brevity, only one simulation set and one AMC procedure were examined to compare the efficiency of variable-length and fixed-length CATs at Time 2. Simulation Set 1 was used because its 50-item fixed-length tests (FLT) allowed for greater potential savings in test length. KL-Z was used because this procedure had performed well in the fixed-length study.

The variable-length test (VLT) 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 FLT, 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 slightly lower power than the FLT) were partially an artifact of the chosen C_1 and C_2. In particular, lowering the critical values C_1 and C_2 would lead to more rejections of H_0. The power of the VLT would thereby be increased—at the expense of the Type I error rate—thus making the classification properties of the VLT and FLT more similar to each other. Conversely, raising the FLT's critical value above 1.96 would lead to fewer rejections of H_0, thereby reducing both the power and Type I error rate and resulting in classification properties more similar to those of the VLT.

Insert Figure 4 about here

Table 2 shows the mean and SD of the number of fewer items administered by the VLT than by the FLT for the three change conditions and the no change condition. Results for the no change condition are shown as a baseline since they represent only Type I error. For the change conditions, mean reductions ranged from 3.70 to 6.55, 19.04 to 21.41, and 28.58 to 29.02 items for change values of 0.5, 1.0, and 1.5, respectively. Thus, the VLT exhibited moderate reductions in the average test length with change of 0.5, and substantial reductions at 1.0 and 1.5. Because

the number of items saved by the VLT is clearly dependent upon the length of the FLT, it is important to also consider the gain in efficiency as a percentage. The mean percentage of items saved ranged from 10% under positive change of 0.5, to 40% under positive change of 1.0, and to over 50% under positive change of 1.5.

Insert Table 2 about here

Table 2 shows that the SD of the number of items saved was small (below 5) under change of 1.5. This finding was due to a "floor" effect, whereby the test lengths tended to cluster around the minimum test length. On the other hand, there was substantial variability in item reductions when the assumed change value was 0.5 or 1.0 (SDs exceeding 9 in all cases). As anticipated, the ratio of the mean savings to the SD of savings increased as the level of true change increased.

Discussion and Conclusions

The objective 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 to detect change than the method based on overlapping confidence intervals. The difference in performance among item selection methods was more subtle, but using Kullback-Leibler information 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. With variable-length testing at Time 2, this

combination was found to yield substantial reductions 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 3 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 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.

Insert Table 3 about here

Table 3 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 3. 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 3 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. Future studies of AMC should investigate its power under different conditions, including varying the item pool characteristics, test lengths, and magnitudes of change. In particular, higher levels of item discrimination should be examined since increased discrimination should decrease standard errors, thus enabling more efficient identification of significant change with variable-length AMC. 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. Further study of variable-length testing, e.g., allowing for variable length at the first time point, stopping early when significant change has not occurred, and using different minimum test lengths.
3. Detection of change at more than two time points.
4. Application of AMC to measurement with polytomous models (e.g., Muraki, 1992; Samejima, 1969), which are often used in psychological assessments.
5. 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, lack of progress, and decline in students, patients, and other applied areas of psychology is an important application of psychometrics that merits further study. The resulting indicators that significant change has or has not occurred can inform treatment decisions, educational interventions, and other applied decisions. 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.

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika, 50,* 3–16.

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.

Bock, R.D. (1976). Basic issues in the measurement of change. In D. N. M. de Gruijter and L. J. T. van der Kamp (Eds.), *Advances in psychological and educational measurement* (pp. 75–96). New York: John Wiley & Sons.

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

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

Collins, L. M., & Sayer, A. G. (Eds.). (2001). *New methods for the analysis of change.* Washington DC: American Psychological Association.

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. (1991a). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56*, 495–515.

Embretson, S. E. (1991b). Implications of a multidimensional latent trait model for measuring change. In L. M. Collins & J. L. Horn (Eds.). *Best methods for the analysis of change* (pp. 184–197). Washington, D.C.: American Psychological Association.

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.

Fischer, G. H. (1976). Some probabilistic models for measuring change. In D. N. M. de Gruijter & L. J. T. van der Kamp (Eds.), *Advances in psychological and educational measurement* (pp. 97–110). New York: John Wiley & Sons.

Hummel-Rossi, B., & Weinberg, S. L. (1975). Practical guidelines in applying current theories to the measurement of change. I. Problems in measuring change and recommended procedures. *JSAS Catalog of Selected Documents in Psychology, 5*, 226. (Ms. No. 916)

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., & 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. (1963). Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 21–38). Madison, WI: The University of Wisconsin Press.

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.

McNemar, Q. (1958). On growth measurement. *Educational and Psychological Measurement, 18*, 47–55.

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.

Tucker, L. R., Damarin, F., & Messick, S. (1966). A base-free measure of change. *Psychometrika, 31*, 457–473.

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. (1994). Measurement of change. In T. Husen & T.N. Postlethwaite (Eds.), *The international encyclopedia of education* (2nd ed., pp. 671–678). Oxford, UK: Pergamon.

Willett, J.B. (1997). Measuring change: What individual growth modeling buys you. In E. Amsel & K. A. Renninger (Eds.), *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.

Acknowledgements

The authors gratefully acknowledge the assistance of Ben Hemingway for providing the data in Table 3. We also thank two anonymous reviewers for their suggested improvements to a previous version of this paper.

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

Time 1 (θ_1)	Time 2 (θ_2)								
	–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

Table 2
Number of Items Saved in Variable-Length Testing Using KL-Z

Change Condition	Time 1 θ								
	–2.0	–1.5	1.0	–0.5	0.0	0.5	1.0	1.5	2.0
No Change									
Mean	0.43	0.49	0.59	0.50	0.48	0.61	0.63	0.60	0.36
SD	3.26	3.55	3.82	3.53	3.42	4.02	4.01	3.92	3.06
θ = 0.5									
Mean	3.70	4.77	6.55	4.69	4.93	5.53	5.35	5.43	--
SD	9.09	10.00	11.39	9.84	10.07	10.49	10.59	10.78	
θ = 1.0									
Mean	19.04	20.01	21.06	20.54	20.78	20.08	21.41	--	--
SD	12.71	12.21	11.85	12.01	12.00	12.12	11.94		
θ = 1.5									
Mean	28.58	28.97	28.78	28.87	29.02	29.00	--	--	--
SD	4.74	4.22	4.71	4.17	3.97	3.81			

Table 3. 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

Figure 1. Fisher Information Function of Each Item Bank

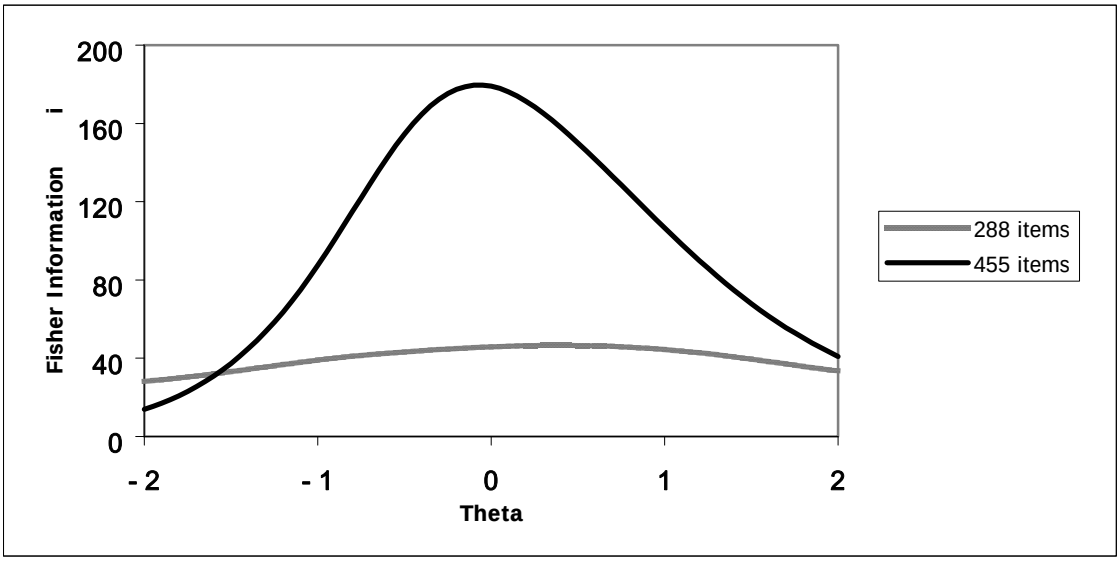

Figure 2. Type I Error Rates and Power Functions, Simulation Set 1

Figure 2a
No Change

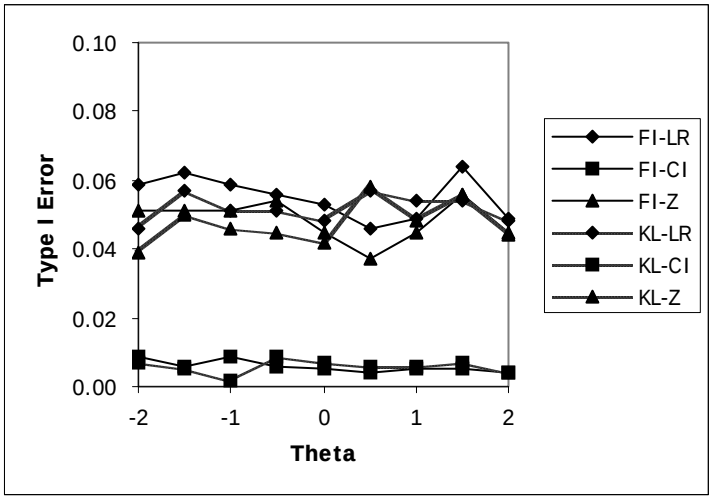

Figure 2b
True Change of 0.5

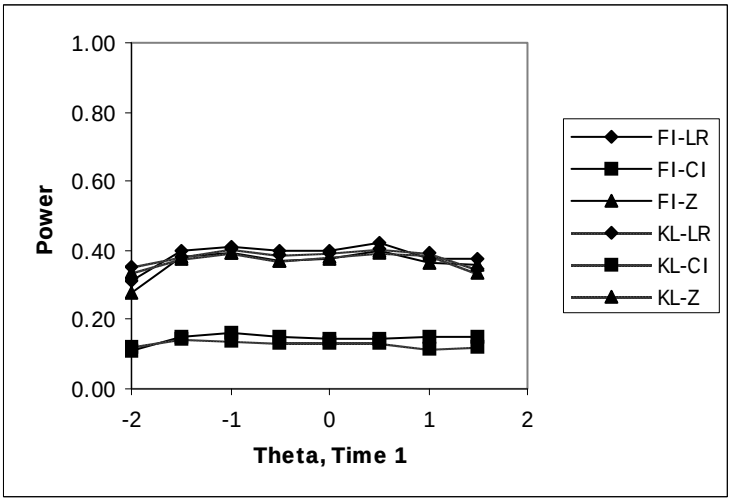

Figure 2c
True Change of 1.0

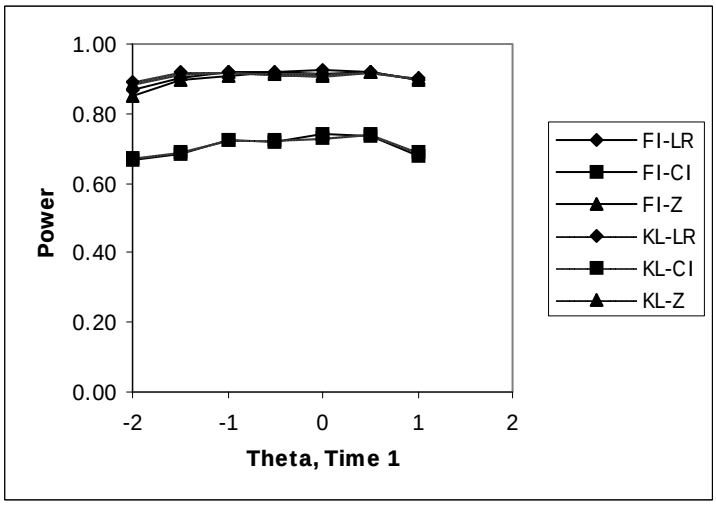

Figure 2d
True Change of 1.5

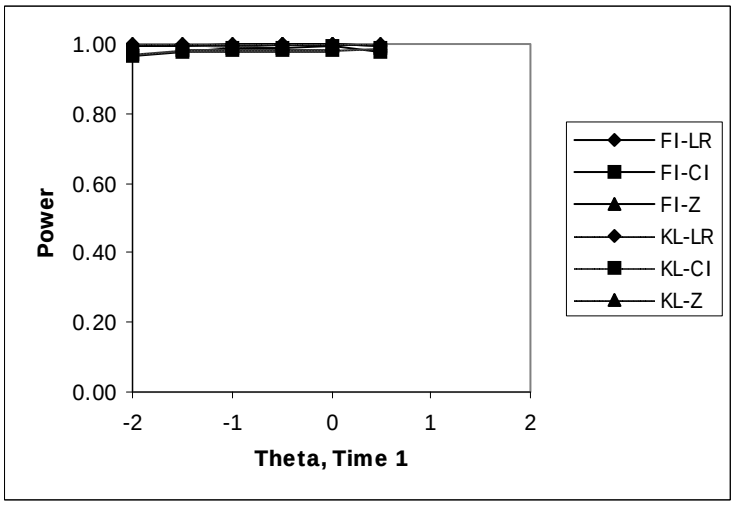

Figure 3. Type I Error Rates and Power Functions, Simulation Set 2

Figure 3a
No Change

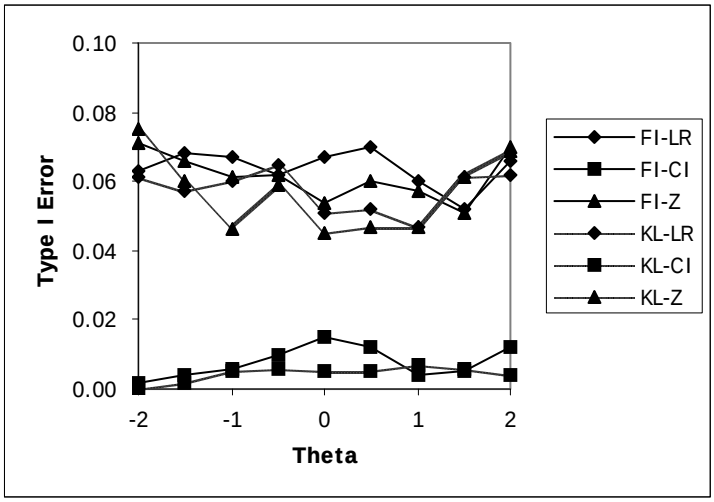

Figure 3b
True Change of 0.5

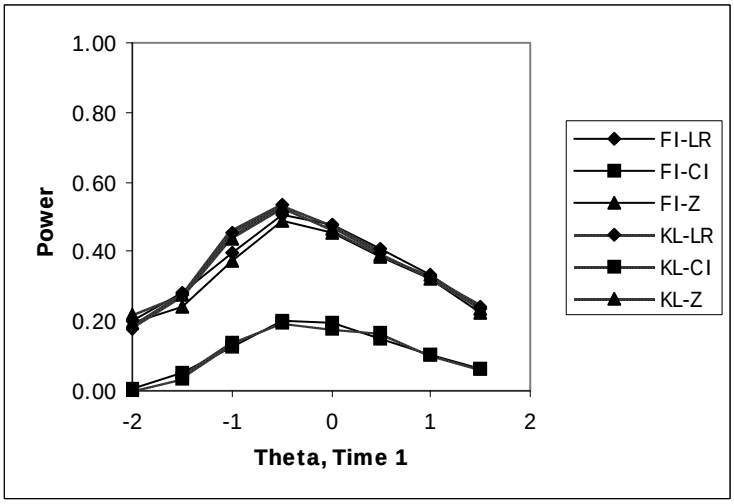

Figure 3c
True Change of 1.0

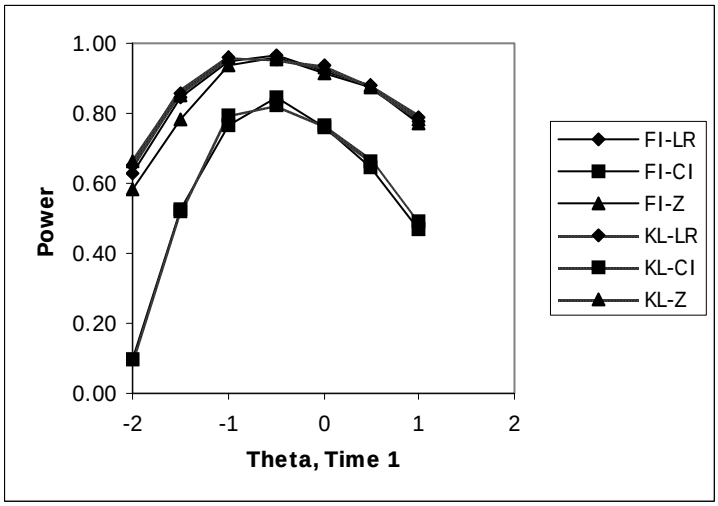

Figure 3d
True Change of 1.5

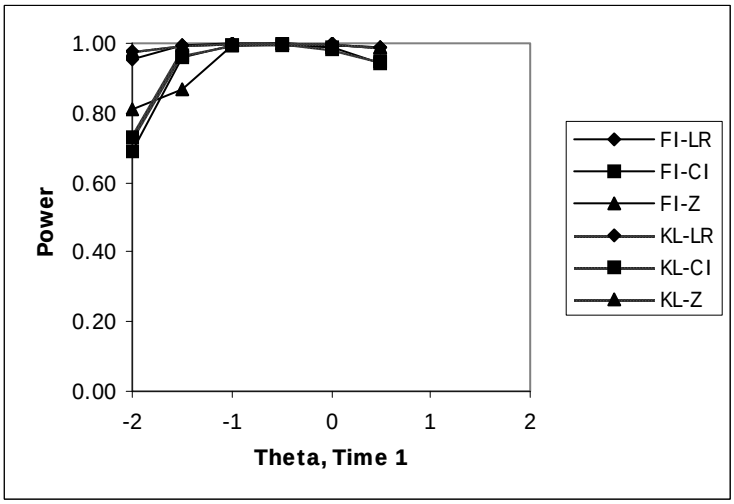

Figure 4: Type I Error Rates and Power Functions for Fixed-Length and Variable-Length Tests Using KL-Z

Figure 4a
No Change

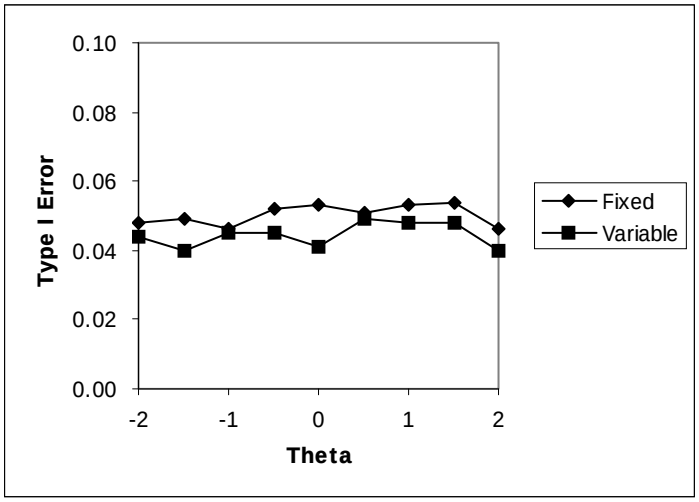

Figure 4b
True Change of 0.5

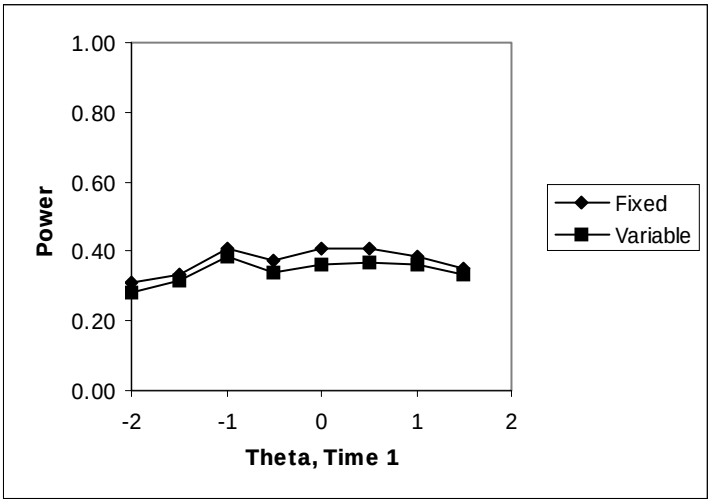

Figure 4c
True Change of 1.0

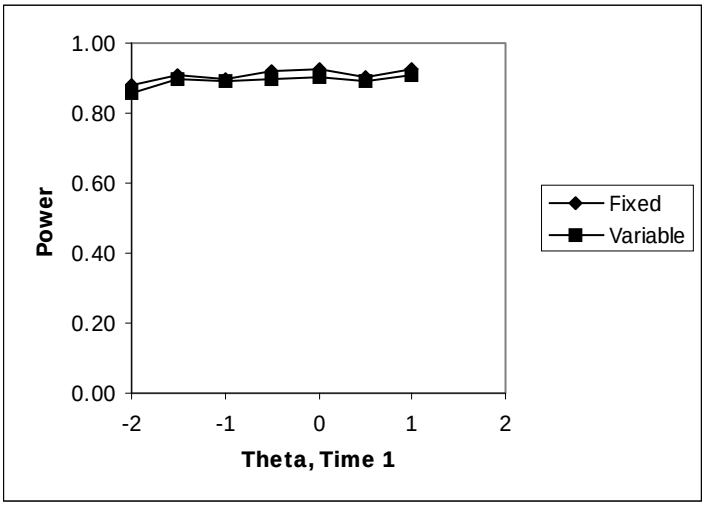

Figure 4d
True Change of 1.5