

Comparison of Computerized Adaptive Testing and Classical Methods for Measuring Individual Change

Gyenam Kim Kang
Korea Nazarene University

David J. Weiss
University of Minnesota

Presented at the Item Calibration and Special Applications Paper Session, June 7, 2007



2007 GMAC® Conference on Computerized Adaptive Testing

Abstract

Monte carlo simulation was used to compare four methods for measuring individual change in terms of recovery of true change: two classical test (CT) methods, one CT-based item response theory (IRT) method, and an adaptive measurement of change (AMC) procedure. The conditions manipulated to evaluate the recovery of true change were (1) the Time 1 trait level, the magnitude and variability of true change at Time 2, and (3) the discrimination of the tests. Results indicated that the CTs used to measure change were heavily affected by the range of θ to which the test was targeted and by the item discrimination level of the test. Scoring of the CTs using IRT methods resulted in improved measurement of change. The AMC method provided the best measurement of change and was able to identify significant individual change with substantially reduced numbers of items.

Acknowledgment

Presentation of this paper at the 2007 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.

Copyright © 2007 by the authors.

All rights reserved. Permission is granted for non-commercial use.

Citation

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*. Retrieved [date] from www.psych.umn.edu/psylabs/CATCentral/

Author Contact

**David J. Weiss, N660 Elliott Hall, University of Minnesota,
Minneapolis MN 55455, U.S.A. Email: djweiss@umn.edu**

Comparison of Computerized Adaptive Testing and Classical Methods for Measuring Individual Change

Educational measurement has been primarily used for purposes of admissions, placement, and evaluating instructional programs at the classroom, district, and state level. However, in a flexible learning environment in which individual differences among students are important, the measurement of individual change could be meaningful in the sense that the goal of education could be defined as each student's change on important educational variables. That is, educational measurement would consider how much a particular student's achievement level has been changed as the result of instructional experiences in a defined curriculum. In this case, educational measurement would be based on continuous measurement of change for a particular student and educational goals would be evaluated with reference to each student's previous achievement level.

The measurement of individual change is also important in clinical applications. Frequently, a client of a counseling program or a patient at a clinic is measured on relevant variables to determine the type of treatment required. For example, a client might present with a high level of depression. Treatment might then be prescribed designed to lower the level of depression. To determine whether the treatment has been successful, the patient's depression is again measured after a period of time and compared to the initial level. If there has been a significant change in the patient's level of depression, treatment might be discontinued. If, however, the patient's depression has not changed significantly or has increased, a different treatment might be prescribed. As in educational measurement, change might be measured at multiple occasions for a single person.

One of the traditional ways to measure individual change is simply to compute the difference between scores obtained at two points of measurement, such as in a pretest-posttest paradigm. This simple difference score is most often given by the formula (Burr & Nesselrode, 1990; McDonald, 1999)

$$D = Y - X \quad (1)$$

where D is the observed change or difference score,
 Y is the observed score at Time 2, and
 X is the observed score at Time 1.

Previous research regarding the simple difference score (SDS) for the measurement of individual change has demonstrated that it has major problems: low reliability (Allen & Yen, 1979; 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) at two or more points of measurement.

Several different procedures for estimating change have been suggested (Lord, 1963; Manning & DuBois, 1962; McNemar, 1958; Traub, 1967; Tucker, Damarin, & Messick, 1966), in addition to the simple difference score. 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, 1989, 1997). Manning and DuBois (1962) showed theoretically that the residual change score is more reliable than the simple difference score in most situations. The

residual change score (R) reflects the difference between an actual and a predicted score; it is computed as

$$R = Y - Y' \tag{2}$$

$$= Y - \bar{Y} - b_{Y.X} (X - \bar{X}) \tag{3}$$

where Y is the observed score at Time 2,

X is the observed score at Time 1,

Y' is the predicted score from X based on the bivariate linear regression of Y on X ,

\bar{X} and \bar{Y} are the means of the distributions of observed scores at Time 1 and Time 2, respectively, and

$b_{Y.X}$ is the slope of the linear regression line for predicting Y from X .

Group level information is required to obtain the RCS. In addition, the RCS 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. Whereas the SCS is typically negatively correlated with the initial scores at Time 1 (Burr & Nesselrode, 1990; Lord, 1963; Hummel-Rossi & Weinberg, 1975; Willett, 1997), the RCS correlates 0.0 with the values at Time 1 since residual values are not correlated with predictor values (Hays, 1988). Therefore, the RCS is appropriate for studying the correlates of change, but not for evaluation of individual change.

Item response theory (IRT) has several advantages over classical test theory (CTT). Scoring in IRT results in scores that can be independent from a specific measuring instrument. With respect to reliability and the standard error of measurement, IRT permits evaluation of an item or a test at each level of a trait (θ). Furthermore, IRT defines the relationship between an individual's performance on the test items and the trait measured by the test as nonlinear. Thus, IRT has the potential to reduce several of the problems inherent in conventional tests that are associated with measuring individual change.

A number of researchers have addressed the issue of measuring change using item response theory (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). However, the model proposed by Embretson is the only IRT model that provides change parameters for measuring individual change in learning, but it is restricted to the one-parameter logistic multidimensional IRT model which requires the unrealistic assumption of equal discriminations across items. The other IRT models estimate group change (Fischer, 1976), require group level information (Bock, 1976), or are not designed to estimate the extent of individual change but to assess the relationship between the latent trait at two time points and/or changes in the latent trait across time (Andersen, 1985).

Although previous research based on CTT (Lord, 1958; Duncan, 1974) and IRT (Embretson, 1991a; 1991b) has provided adequate means of measuring change in some situations, each of the CTT and IRT approaches to date is limited. It is apparent that a different approach is required to measure individual change more accurately.

Adaptive Measurement of Change

Weiss and Kingsbury (1984) proposed a method they called adaptive self-referenced testing as a method for measuring individual change by taking advantage of the benefits of both IRT and

computerized adaptive testing. This procedure is referred to here as the adaptive measurement of change (AMC). The characteristics of adaptive tests are that different items, or sets of items, are administered to different individuals depending on each individual's status on the latent trait (Hambleton, Swaminathan, & Rogers, 1991; Weiss, 1982, 1983, 1985, 2004; Weiss & Kingsbury, 1984). Adaptive testing provides the opportunity to match an individual's trait level with item difficulty, and the most informative test can be administered to each individual (Hambleton & Swaminathan, 1985; Weiss, 1995). Adaptive testing has been expanded over the years to incorporate IRT procedures and computer administration of the test, and is known as computerized adaptive test (CAT).

AMC uses CAT and IRT to obtain estimates of an individual's θ level from a domain of items on occasions separated by an interval of time, and the measurement of change for a particular student is determined with reference to that student's previous θ level. AMC incorporates the six components that define a CAT: (1) an item response model, (2) a starting trait estimate for each examinee, (3) a pre-calibrated item bank, (4) an item scoring procedure, (5) an item selection rule, and (6) a termination criterion (Weiss & Kingsbury, 1984).

In AMC, change is measured as the difference between estimated θ levels ($\hat{\theta}$) for two (or more) occasions. Significant change is said to occur, and the Time 2 test is terminated, when the IRT-based confidence intervals for two $\hat{\theta}$ s do not overlap (Weiss & Kingsbury, 1984). The confidence intervals, or standard error (SE) bands, are generally calculated as:

$$\hat{\theta} \pm 2 SEM | \hat{\theta}_j. \quad (4)$$

where $SEM | \hat{\theta}_j$, the reciprocal square root of response pattern information $\left[1/\sqrt{I(\hat{\theta}_j)}\right]$, is determined from the second derivative of the log-likelihood function (Assessment Systems Corporation, 2005),

$$SEM | \hat{\theta}_j = \sqrt{Var(\hat{\theta}_j | \theta)}, \quad (5)$$

where

$$Var(\hat{\theta}_j | \theta) = \frac{1}{I(\hat{\theta}_j)} \quad (6)$$

and

$$I(\hat{\theta}_j) = \left(\frac{\partial^2 \ln L}{\partial \theta_j^2} \right), \quad (7)$$

with θ_j evaluated at $\hat{\theta}_j$.

The measurement of change for an individual by AMC is determined with reference to each individual with no reference to data from other individuals. Therefore, the measurements reflect how the individual's θ estimate at Time 2 differs from their own θ level at Time 1. However, there has been little empirical research involving AMC for measuring change at the individual level.

The objectives of this study were to compare the feasibility of AMC with CTT methods in measuring individual change, by examining the recovery of true change and by identifying particular experimental conditions under which each procedure better recovered true change.

Method

Monte carlo simulation was used to compare four methods for measuring individual change in terms of recovery of true change: two CT methods, one CT-based IRT approach, and AMC. The conditions manipulated to evaluate the recovery of true change were (1) the Time 1 (T1) trait level, (2) the magnitude and variability of true change at Time 2 (T2), and (3) the discrimination of the tests.

True θ Distributions

The T1 true θ values of 1,500 simulated examinees were generated to have a rectangular distribution with mean 0.0 and standard deviation (SD) of 1.3 and a range from -2.25 to $+2.25$. The true θ range at T1 was divided into three groups—low (-2.25 to -0.75), medium ($-.7499$ to $+0.7499$), and high ($+0.75$ to $+2.25$)—to evaluate the results conditional on θ .

The distribution of true θ at T2 for the 1,500 examinees was generated to reflect only positive true change. For each of the three initial θ groups, nine different true change conditions were formed: three different levels of average true change (low mean: 0.5, medium mean: 1.0, and high mean: 1.5) were crossed with three different levels of variability of true change (low SD: .01, medium SD: .05, and high SD: .10). The nine different true change conditions involving the magnitudes and variabilities of true change, were abbreviated as follows: LL (0.5, 0.01), LM (0.5, 0.05), LH (0.5, 0.10), ML (1.0, 0.01), MM (1.0, 0.05), MH (1.0, 0.10), HL (1.5, 0.01), HM (1.5, 0.05), and HH (1.5, 0.10). True θ values at T2 were obtained by adding the corresponding true change to the true θ value at T1. As a result, 27 true change conditions for T2 were formed across the entire range of initial true θ in each item discrimination condition.

Item Banks

Three item banks were generated to have different average item discrimination conditions using the 3-parameter logistic model

$$P_{ij}(\theta_j) = c_i + (1 - c_i) \frac{\exp[Da_i(\theta_j - b)]}{1 + \exp[Da_i(\theta_j - b)]}, \quad (8)$$

where P_{ij} is the probability of a correct response to item i by simulee j ,

θ_j is the trait level for simulee j ,

a_i is the discrimination parameter for item i ,

b_i is the difficulty parameter for item i ,

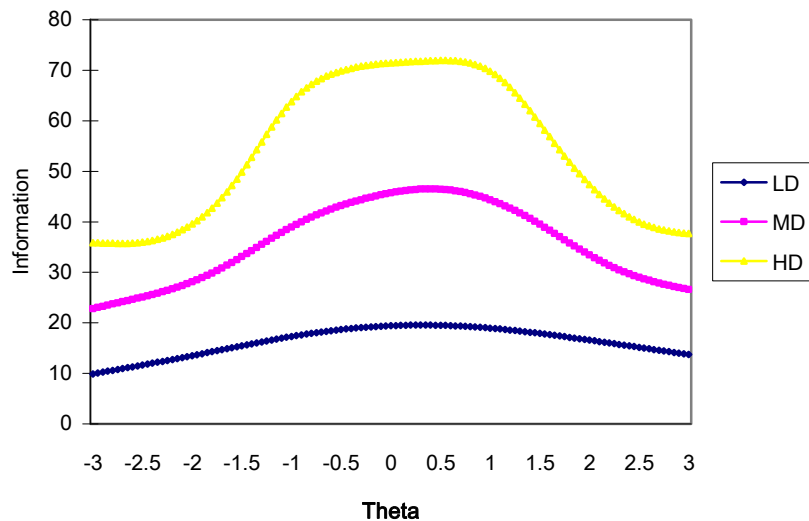
c_i is the pseudo-guessing parameter for item i , and

$D = 1.7$.

Three different average item discrimination conditions were low (LD; $\bar{a} = 0.5$), medium (MD; $\bar{a} = 1.0$), and high (HD; $\bar{a} = 1.5$), respectively, with average standard deviation (SD) of 0.15 in each item bank. However, the distribution of item difficulties was peaked rectangular, indicating that a sufficient number of items would be available in the middle range for

conventional and adaptive tests from the same item bank. The distribution of item difficulties was centered at 0.00 and had 18 intervals from -4.50 to $+4.50$, with a range wider than the true T1 θ range $[-2.25$ to $+2.25]$. The middle six intervals, -1.5 to $+1.5$, contained 24 items, while the other twelve intervals (the six low and six high intervals) contained only 12 items per interval, so each item bank consisted of 288 items, peaked at the middle intervals. The pseudo-guessing parameter for all the items was fixed at .20 for all test conditions (Kingsbury & Weiss, 1983; Lord & Novick, 1968; Urry, 1977; Yen, 1986). Figure 1 shows the bank information functions for the three item banks.

Figure 1
Bank Information Functions for the LD, MD, and HD Item Banks



Conventional Tests

The conventional tests (CTs) were constructed as parallel tests to measure individual change (Friedenberg, 1995). From each item bank, two parallel 50-item fixed-length CTs were constructed as peaked tests to measure well at a specific level of difficulty (Tinkelman, 1971; Weiss, 1985). In each of the three item banks, items for the first CT were selected at random from the middle six intervals, containing 24 items in each interval in the corresponding item bank, ranging from -1.5 to $+1.5$ with average item difficulty level of 0.0. Items for the parallel tests were selected at random from those items not previously selected in each item bank. Six CTs were constructed from the three different item banks.

The item responses for each of the six CTs were generated using the values of true θ_1 (or θ_2) and the item parameters of each set of 50 items using PARDSIM (Assessment Systems Corporation, 1997). The probability of a correct response to each item in the test for each simulee was generated in accordance with the 3-parameter logistic model. Then the model-generated probability matrix was converted to a 1-0 score data matrix by comparing cell-by-cell with a matrix of random numbers generated from a rectangular distribution.

The number correct (NC) scores and three different estimates of θ —maximum likelihood (ML) estimates ($\hat{\theta}_{C1}$ and $\hat{\theta}_{C2}$ for ML), the Bayesian modal (MAP) estimates ($\hat{\theta}_{C1}$ and $\hat{\theta}_{C2}$ for

MAP) and the expected a posteriori (EAP) Bayesian estimates ($\hat{\theta}_{C1}$ and $\hat{\theta}_{C2}$ for EAP)—were obtained using SCOREALL (Assessment Systems Corporation, 1998). The NC scores were also transformed to the θ metric ($\hat{\theta}_{NC1}$ and $\hat{\theta}_{NC2}$) using the test response function (TRF) to enable a direct comparison between estimates from CTs and those from AMC.

CATs

The item responses of all simulees in each of the three item banks were generated using the values of true θ_1 (or θ_2) and item characteristics for the 288 items using PARDSIM (Assessment Systems Corporation, 1997). The procedure to generate the item responses was the same as for the CTs, only the number of items included was different (50 for the CTs and 288 for the CATs). The model-generated probability matrix was constructed and converted to a 1-0 score data matrix by comparing a matrix of rectangular random numbers. As a result, a $1,500 \times 288$ item response matrix was constructed for CAT.

The administration of CAT was performed using POSTSIM, a program for post-hoc simulation of CAT (Assessment Systems Corporation, 2005). In administering CAT, the initial θ value was the same, $\theta = 0$ for all simulees at T1. ML estimation of θ (Yoes, 1993; Zwick, Thayer, & Wingersky, 1994) was used at T1 ($\hat{\theta}_{A1}$) and T2 ($\hat{\theta}_{A2}$). However, ML estimation cannot be used for non-mixed response patterns, so a step size of ± 3 was used to select the next item to obtain mixed response patterns as early in the CAT as possible. Items in the CATs were selected to provide maximum information at each $\hat{\theta}$ in the CAT (Hambleton & Swaminathan, 1985; Weiss, 1982; Weiss & Kingsbury, 1984). Finally, the CAT procedure was terminated after the administration of 50 items at both T1 and T2 to make direct comparisons by matching the number of items of CTs. The final MLE of θ obtained at T1 ($\hat{\theta}_{A1}$) was used as the T2 CAT entry level to make the CAT more efficient.

Four Approaches to Measuring Change

SDS. The SDS, defined in Equation 1, was obtained as the difference between the TRF-transformed value of NC to the θ metric at T1 and T2, such that

$$SDS = \hat{\theta}_{NC2} - \hat{\theta}_{NC1}, \quad (9)$$

where, $\hat{\theta}_{NC2}$ is the transformed value of NC scores to the θ metric at T2, and

$\hat{\theta}_{NC1}$ is the transformed value NC scores to the θ metric at T1.

RCS. The RCS was defined in Equation 2 as the difference between the observed score at T2 and the predicted value based on the bivariate linear regression of observed T2 status on observed initial status. The RCS was expressed using TRF-transformed values of NC scores to the θ metric,

$$\begin{aligned} RCS &= \hat{\theta}_{NC2} - \hat{\theta}_{NC2} \\ &= \hat{\theta}_{NC2} - \bar{\theta}_{NC2} - b_{\hat{\theta}_{NC2}, \hat{\theta}_{NC1}} \left(\hat{\theta}_{NC1} - \bar{\theta}_{NC1} \right) \end{aligned} \quad (10)$$

where $\hat{\theta}_{NC2}$ and $\hat{\theta}_{NC1}$ are the TRF-transformed values corresponding to the NC scores (Y and X)

at Time 2 and Time 1, respectively, and

$$b_{\hat{\theta}_{NC2}, \hat{\theta}_{NC1}} = r_{\hat{\theta}_{NC2}, \hat{\theta}_{NC1}} \left(\frac{s_{\hat{\theta}_{NC2}}}{s_{\hat{\theta}_{NC1}}} \right) \quad (11)$$

is the TRF-transformed value of $b_{Y.X}$

IRT-scored difference score. The IRT-scored difference score (*IRTDS*) was defined as the difference between θ estimates ($\hat{\theta}_{C1}$ and $\hat{\theta}_{C2}$) from the conventional test at T1 and T2. The estimates of the underlying trait, $\hat{\theta}_{C1}$ and $\hat{\theta}_{C2}$, were not the transformed value of NC, but were directly obtained from the data of the conventional tests using SCOREALL. The following IRTDS was obtained for each of three θ estimates—ML, MAP, and EAP,

$$IRTDS = \hat{\theta}_{C2} - \hat{\theta}_{C1}, \quad (12)$$

where, $\hat{\theta}_{C2}$ is the θ estimate at T2 from CT,

$\hat{\theta}_{C1}$ is the θ at T1 from CT.

AMC difference score. The AMC difference score (AMCDS) was defined as

$$AMCDS = \hat{\theta}_{A2} - \hat{\theta}_{A1}, \quad (13)$$

where, $\hat{\theta}_{A1}$ is the examinee's θ estimate at T1 from AMC, and

$\hat{\theta}_{A2}$ is the AMC θ estimate at T2.

Evaluation of the Procedures for Measuring Change

Evaluation criteria. How well each of the four approaches recovered true change was evaluated using three criteria—Pearson product-moment correlation coefficients, root mean square error (RMSE), and the average bias between true and estimated change values—for each of the nine different change conditions in each initial (T1) θ level within each of the three different item discrimination conditions.

The Pearson product-moment correlation coefficient was computed as $r_{d, \hat{d}}$, where, d is the true change value ($\theta_2 - \theta_1$) and \hat{d} is each of the observed change values (SDS, RCS, IRTDS, or AMCDS).

Root mean square error (RMSE) was computed by

$$RMSE = \sqrt{\frac{\sum_{j=1}^N (d_j - \hat{d}_j)^2}{N}}, \quad (14)$$

Since RMSE is an index that indicates error, values close to 0.0 indicate error-free estimates.

Average bias indicates whether estimation error is systematic and was computed as:

$$\text{bias} = \frac{\sum_{j=1}^N (d_j - \hat{d}_j)}{N}, \quad (15)$$

Positive bias indicates under-estimating the true change and a negative value reflects over-estimating change.

Effect size. Effect sizes were computed to identify influential condition(s) in the recovery of true change. They were obtained from a repeated measures ANOVA design for each of the three evaluative criteria to assist in the evaluation of the estimation procedures. The ANOVA was conducted with two between-subjects factors [item discrimination test conditions (LD, MD, and HD)] and three levels of T1 θ (θ_1 : Low, Medium, and High), and three within-subjects factors—approaches to measuring change (SDS, RCS, IRTDS, and AMCDS), three levels of magnitude of true change, and three levels of variability of true change—for each of the three evaluative criteria (Howell, 1992).

Effect size was calculated as

$$\eta^2 = \frac{SS_{\text{Effect}}}{SS_{\text{total}}}, \quad (16)$$

where, η^2 is the effect size,

SS_{Effect} is the sum of square of each main effect or interaction, and

SS_{total} is the total sum of squares.

Because the distributions of r , RMSE and bias were skewed, they were transformed as follows (Hays, 1988; Howell, 1992; Yoes, 1993):

$$r_z = \frac{1}{2} \cdot \ln\left(\frac{1+r}{1-r}\right), \quad (17)$$

$$\text{LMSE} = \log_{10}(\text{RMSE} + 1), \quad (18)$$

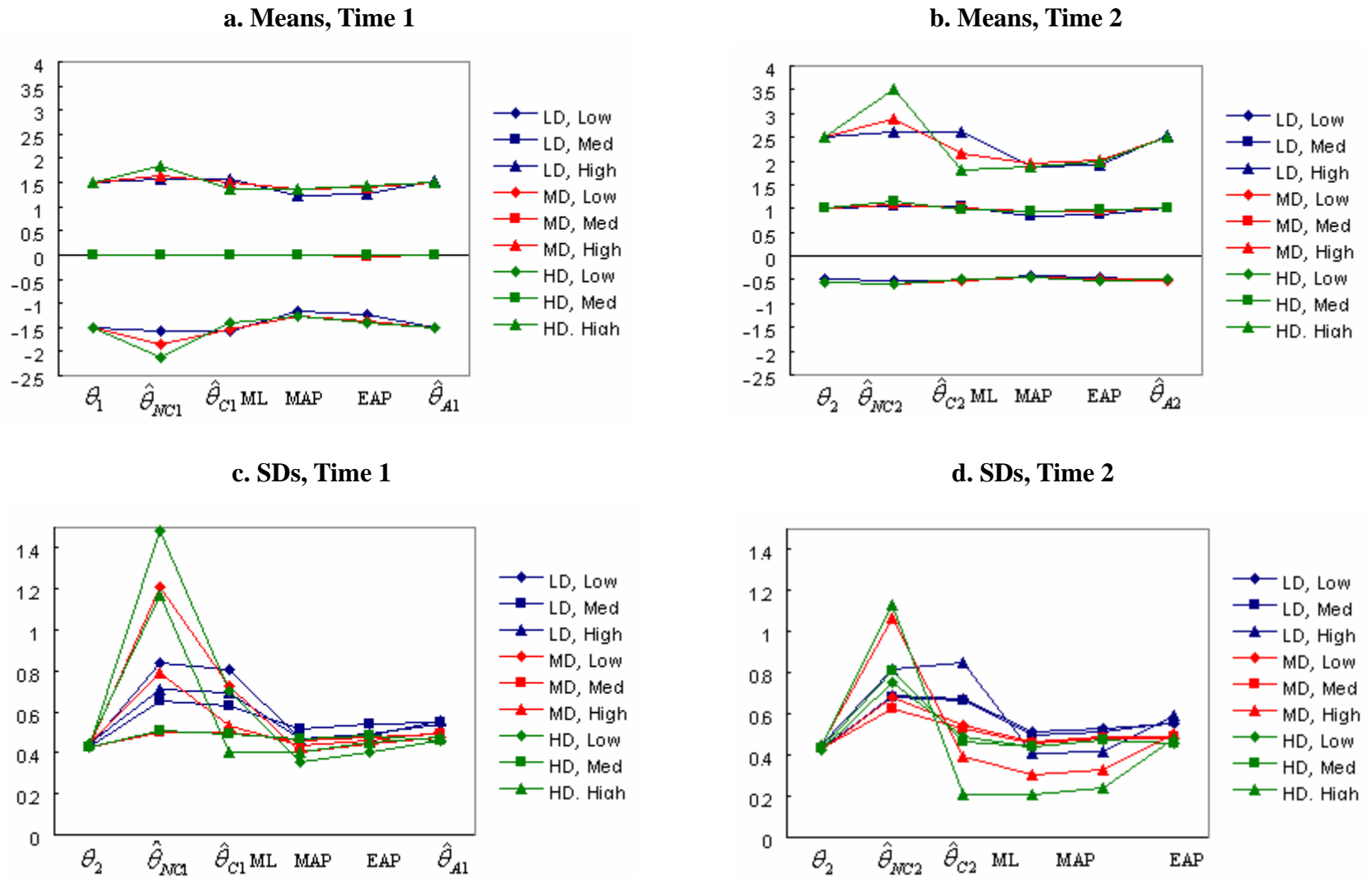
$$\text{LBIAS} = \log_{10}(\text{bias} + 1). \quad (19)$$

Results

Descriptive Statistics

Figure 2 show the means and SDs of true θ , estimated θ , and NC scores for all change conditions and for all item discrimination conditions. The mean $\hat{\theta}_{NC1}$ and $\hat{\theta}_{NC2}$ values were smaller than the corresponding mean true T1 θ (θ_1) and T2 θ (θ_2), respectively, for the low T1 θ group (Figures 2a and 2b), while they were noticeably larger than the corresponding θ_1 and θ_2 for the T1 high θ group. This trend was pronounced for the MD and HD test condition. Since the CTs were targeted to the medium T1 θ level, lower estimates at T1 reflect the fact that the CT was targeted to a level above this group, and higher estimates reflect the fact that the CT was targeted to a level below this group.

Figure 2
Means and SDs of True and Estimated Time 1 and Time 2 Scores for the LD, MD, and HD Conditions



The mean ML $\hat{\theta}_{C1}$ and $\hat{\theta}_{C2}$ reflected the corresponding true T1 θ and T2 θ values, respectively, for the LD test condition; however, they were noticeably smaller for the high T1 θ level of the HD test condition. The mean MAP and EAP $\hat{\theta}_{C1}$ and $\hat{\theta}_{C2}$ values were even lower than the ML $\hat{\theta}$ s for the high T1 θ group and larger than the ML $\hat{\theta}$ s for the low T1 θ group as the discrimination of the test items increased. In the process of obtaining the ML $\hat{\theta}$ s, the perfect score cases were removed and the N varied across test conditions, becoming smaller as the discrimination of the test items and T1 θ level increased.

The observed mean $\hat{\theta}_{A1}$ and $\hat{\theta}_{A2}$ reflected the corresponding true θ values (θ_1, θ_2) across all conditions. The SDs of $\hat{\theta}_{A1}$ and $\hat{\theta}_{A2}$ were also very similar across all conditions (Figures 2c and 2d).

Correlations Between T1 and T2 $\hat{\theta}$ s and θ s

Figure 3 shows the correlations between T1 and T2 $\hat{\theta}$ s, and between the $\hat{\theta}$ s and the θ s. The correlations based on CTs were higher for the medium T1 θ group than for the low and high T1 θ group across all item discrimination test conditions, since the CT was targeted to the medium T1 θ group. The correlations based on CTs increased as the discrimination of the test items increased.

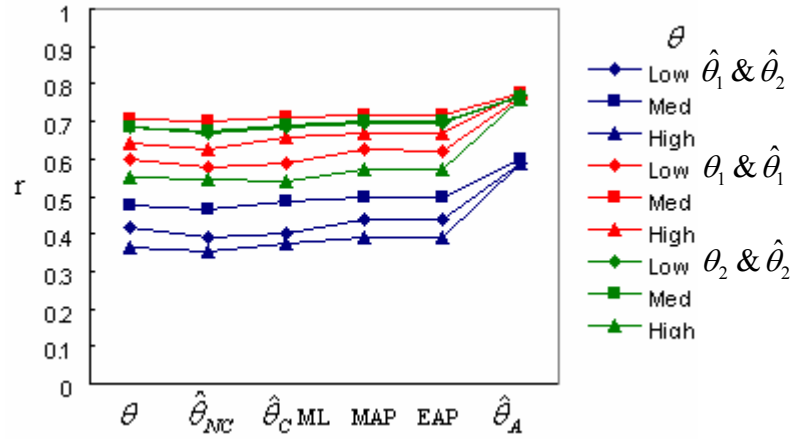
In comparing correlations between θ and the NC scores (θ_1 and NC_1 , θ_2 and NC_2), between the true θ and the NC score transformed to the θ metric (θ_1 and $\hat{\theta}_{NC1}$, θ_2 and $\hat{\theta}_{NC2}$), and between the true and estimated θ (θ_1 and $\hat{\theta}_{C1}$, θ_2 and $\hat{\theta}_{C2}$), the correlations between θ_1 and $\hat{\theta}_{NC1}$, and between θ_2 and $\hat{\theta}_{NC2}$ were the lowest and the correlations between θ_1 and $\hat{\theta}_{C1}$, and between θ_2 and $\hat{\theta}_{C2}$ by MAP and EAP were the highest for low and medium T1 θ levels under all item discrimination test conditions. Under some conditions, the correlations between θ_1 and $\hat{\theta}_{C1}$, and between θ_2 and $\hat{\theta}_{C2}$ by ML, were larger than those between θ_1 and NC_1 , and between θ_2 and NC_2 , respectively; however, the differences in these correlation coefficients were not substantial.

Among the three correlations involving $\hat{\theta}_{C1}$ and $\hat{\theta}_{C2}$ by ML, MAP, and EAP based on CTs, the correlations involving the ML $\hat{\theta}$ s were noticeably smaller than those by MAP and EAP for the MD and HD test conditions. The correlations by MAP and EAP were very similar to each other across all conditions.

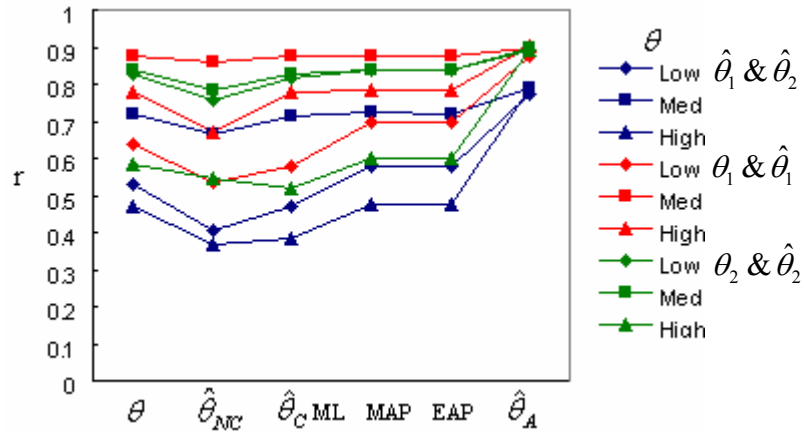
The correlations involving the estimates from CAT ($\hat{\theta}_{A1}$ and $\hat{\theta}_{A2}$, θ_1 and $\hat{\theta}_{A1}$, and θ_2 and $\hat{\theta}_{A2}$) were the highest among the approaches examined and increased as the discrimination of test items increased. Furthermore, unlike the correlation results from CTs, those from adaptive testing were similar across the true T1 θ levels.

Figure 3
Correlations Between T1 and T2 $\hat{\theta}_s$, and Between θ and T1 and T2 $\hat{\theta}_s$

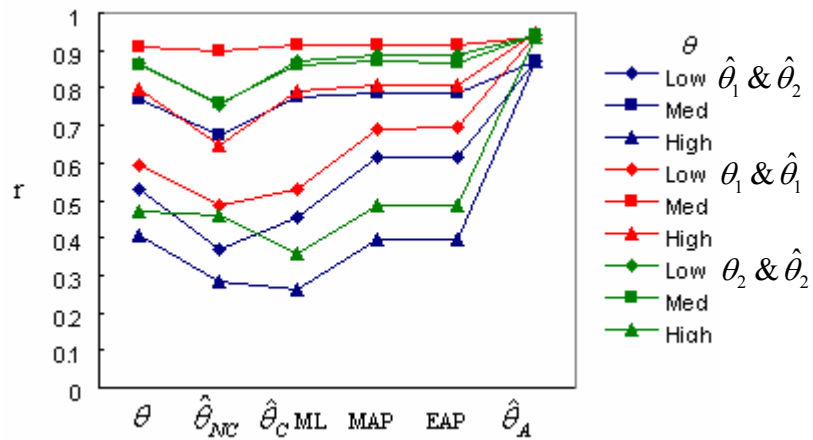
a. LD Condition



b. MD Condition



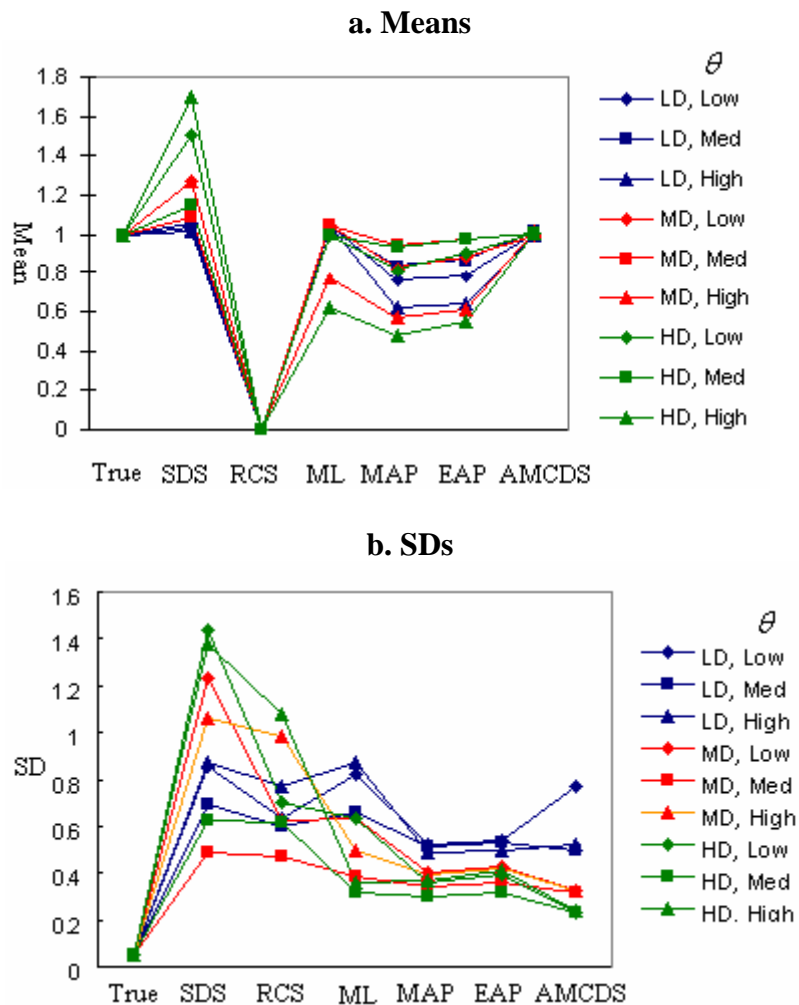
c. HD Condition



Change Scores

Means and SDs. Figure 4 shows the means and SDs of change scores. The mean AMCDS values reflected the average true change values across all conditions (Figure 4a). The SDs of AMCDS were, in general, smaller than those from CTs (SDS, RCS, IRTDS) for most T1 θ levels and for all item discrimination test conditions (Figure 4b). Furthermore, Figure 4b also shows that the SDs of AMCDS decreased as the discrimination of test items increased; however, they were still noticeably larger than those of true change.

Figure 4
Means and SDs of Change Scores for LD, MD, HD Conditions



The mean change scores from CTs (SDS, RCS, IRTDS) were somewhat different. For the LD test condition, the SDS and ML IRTDS reflected the average true change values, in general. However, the MAP and EAP IRTDS reflected less than the average true change values for all T1 θ levels. For the MD and HD condition, the SDS reflected more than the average true change for low and high T1 θ levels (Figure 4a). All three IRTDS results (for ML, MAP, and EAP)

reflected less than the average true change values for high T1 θ under the MD and HD condition. The ML IRTDS values reflected the average true change values for the low and medium T1 θ level; however, those by MAP and EAP reflected the average true change values for the medium T1 θ level, but less than the average true values for the high T1 θ level.

The SDs of the observed change scores were substantially larger than the corresponding SDs of the true change scores (Figure 4b). The SDs of the SDS values were the largest and those of the AMCDS were, in general, the smallest among all the observed change scores across all conditions. The SDs of the SDS values and ML IRTDS results were noticeably smaller for the medium T1 θ level than for the low and high T1 θ levels. Those of MAP and EAP IRTDS and AMCDS were relatively similar across T1 θ levels. The SDs of all observed change scores decreased, in general, as the discrimination of test items increased.

In general, the change scores based on CTs reflected the average true change values for the medium T1 θ level, but reflected more or less than the average true values for the low and high T1 θ levels. In comparing the observed change scores, it can be concluded that the AMC best reflected the average true change values and resulted in the smallest SDs across all conditions examined in this study.

Correlations between observed change score scores and true and estimated θ at T1. The correlations between the observed change scores and the true T1 θ , and the observed change scores and the T1 $\hat{\theta}$ are presented in Figure 5. The average correlations between most of the observed change scores based on CTs (SDS, IRTDS by ML, MAP and EAP) and the true θ at T1 (θ_1) were near 0.0 for the LD test condition (Figure 5a). For the MD and HD test condition, they were noticeably larger than 0.0 in either the positive or negative direction. AMC was the only method for measuring individual change for which estimated change scores were correlated 0.0 with true initial θ across all conditions.

The mean correlations between most of the observed change scores (SDS, IRTDS, and AMCDS) and the estimated θ at T1 ($\hat{\theta}_{NC1}$, $\hat{\theta}_{CI}$, and $\hat{\theta}_{AI}$) were substantially smaller than 0.0 across all conditions (Figure 5b).

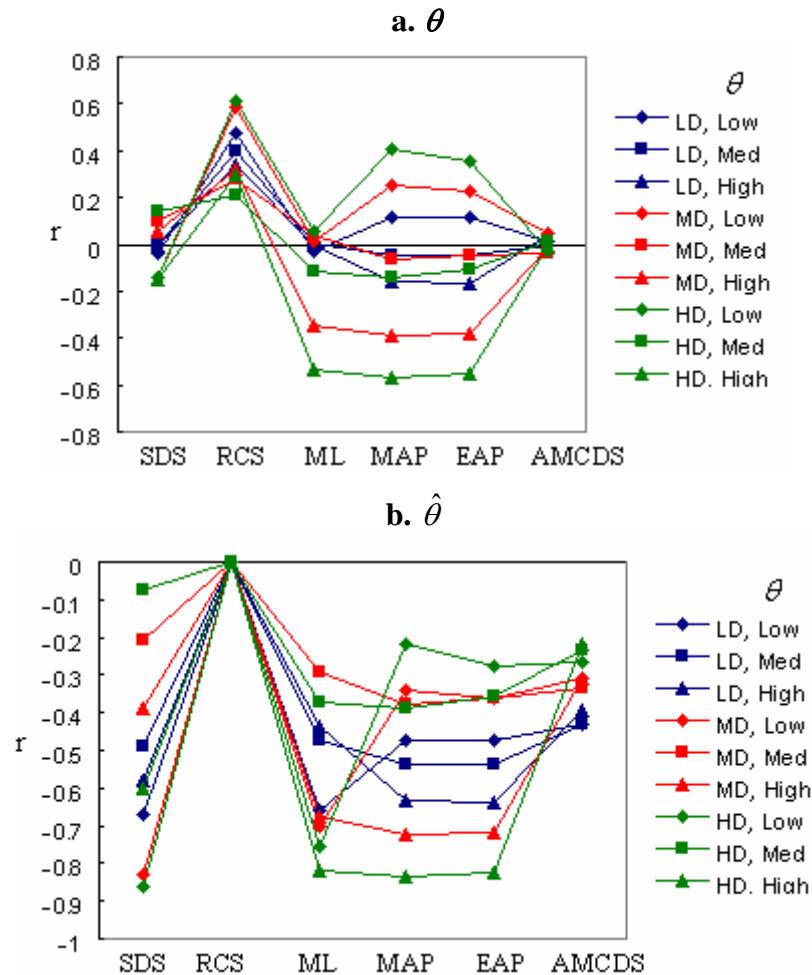
The average correlations between the AMCDS and the estimated θ at T1 ($\hat{\theta}_{AI}$) were the smallest, in general, among the observed change scores examined, and increased in the negative direction as the discrimination of test items increased. In addition, the SDs of AMCDS were similar to each other across T1 θ levels in each of the three different item discrimination test conditions.

The average correlations between RCS and the estimated θ at T1 ($\hat{\theta}_{NC1}$) were exactly 0.0, as expected, and those between RCS and the true θ at T1 (θ_1) were substantially greater than 0.0 across all conditions, ranging from .117 to .688. Since RCS was not correlated with the estimated θ at T1 they can be appropriate for studying correlates of change, but not for evaluation of individual change.

Recovery of True Change

Pearson correlations. The Pearson product-moment correlations as the index for recovery of true change by the observed change scores are provided in Figure 6. AMCDS had consistently

Figure 5
Correlations of T1 θ and $\hat{\theta}$ With Estimated Change Scores
for LD, MD, and HD Conditions



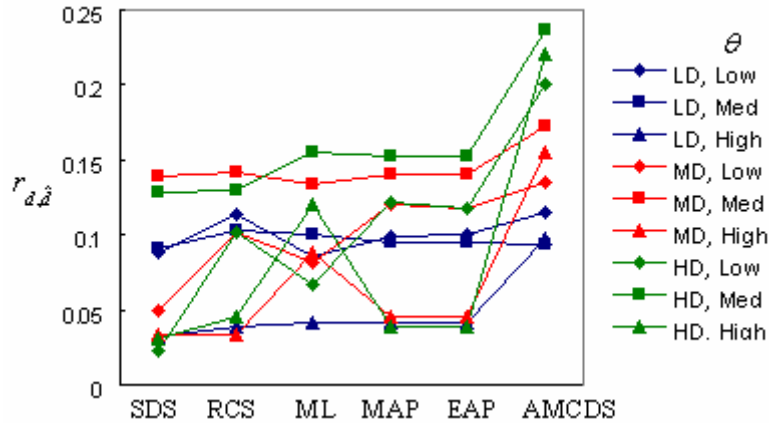
highest correlations of true change with estimated change across all conditions (Figure 6). The results of the repeated-measures ANOVA [sum of squares (SS) and η^2] calculated for the transformed r coefficients are presented in Table 1. The largest effect size for the transformed r coefficients was due to the variability of true change, which accounted for 45% of the total variance in the recovery of true change. The average r values for low, medium, and high levels of variability of true change, across all item discrimination test conditions, T1 θ levels, magnitudes of true change, approaches to measuring change, were .020, .097, and .185, respectively. These results show that as the true change was more variable, the more likely it was to be recovered by the observed change scores.

The effect having the second largest effect size for the transformed r coefficients was due to the approach to measuring change, which accounted for 11% of the total variance in the recovery of true change. The average r values for the SDS, RCS, IRTDS by ML, IRTDS by MAP, IRTDS by EAP, and AMCDS across all other conditions were .068, .090, .097, .095, .094 and .159, respectively, indicating that the AMCDS best recovered true change among all the observed change scores examined, followed by IRTDS, RCS, and SDS.

Table 1
Sum of Squares (SS) and η^2 From the Repeated Measures ANOVA for the Recovery of True Change,

Source of Variation	<i>r</i>		RMSE		Bias	
	SS	η^2	SS	η^2	SS	η^2
Between Subjects Effects						
Level of θ at Time 1 (A)	0.256	0.071	0.246	0.064	0.028	0.002
Test Condition (B)	0.077	0.021	0.017	0.004	0.346	0.027
AB	0.043	0.012	0.052	0.014	0.133	0.010
Within Subjects Effects						
Approach to Measuring Change (C)	0.403	0.112	2.587	0.674	8.321	0.656
AC	0.097	0.027	0.178	0.046	0.705	0.056
BC	0.149	0.041	0.226	0.059	1.033	0.081
ABC	0.015	0.004	0.048	0.013	0.560	0.044
Magnitude of True Change (D)	0.005	0.001	0.171	0.045	0.104	0.008
AD	0.092	0.026	0.034	0.009	0.010	0.001
BD	0.012	0.003	0.002	0.001	0.055	0.004
ABD	0.048	0.013	0.013	0.003	0.025	0.002
Variability of True Change (E)	1.629	0.451	0.000	0.000	0.000	0.000
AE	0.065	0.018	0.000	0.000	0.000	0.000
BE	0.035	0.010	0.000	0.000	0.000	0.000
ABE	0.027	0.008	0.000	0.000	0.000	0.000
CD	0.014	0.004	0.157	0.041	0.953	0.075
ACD	0.033	0.009	0.065	0.017	0.120	0.009
BCD	0.008	0.002	0.011	0.003	0.070	0.014
ABCD	0.019	0.005	0.022	0.006	0.102	0.008
CE	0.203	0.056	0.000	0.000	0.000	0.000
ACE	0.023	0.006	0.001	0.000	0.001	0.000
BCE	0.059	0.016	0.000	0.000	0.000	0.000
ABCE	0.029	0.008	0.000	0.000	0.002	0.000
DE	0.005	0.001	0.000	0.001	0.000	0.000
ADE	0.070	0.019	0.001	0.000	0.001	0.000
BDD	0.013	0.003	0.000	0.000	0.002	0.000
ABDE	0.072	0.020	0.001	0.000	0.003	0.000
CDE	0.010	0.003	0.000	0.000	0.001	0.000
ACDE	0.036	0.010	0.001	0.000	0.036	0.003
BCDE	0.009	0.003	0.001	0.000	0.005	0.000
ABCDE	0.053	0.015	0.003	0.001	0.000	0.000
Total	3.609		3.837		12.693	

Figure 6
Recovery of True Change as Indexed by r



Although Figure 6 shows that the AMCDS consistently had higher correlations of estimated change with true change, the correlations were not high, ranging from about .10 to .24 for AMCDS. These correlations, however, were affected by restriction of the range of true change within each of the θ conditions. Table 2 displays correlations in the recovery of change for each method of change combined across all nine combinations of θ levels and variability of true change. As the table shows, correlations for the AMCDS ranged from .637 for the LD condition to .878 for the HD condition. Within each of the item discrimination conditions, AMCDS had the highest correlations. An interesting comparison is between the SDS and the IRTDSs. These results show that simply scoring the same item responses from CTs using IRT θ estimation methods improves the recovery of true change, with ML on average showing the most improvement of the three θ estimation methods.

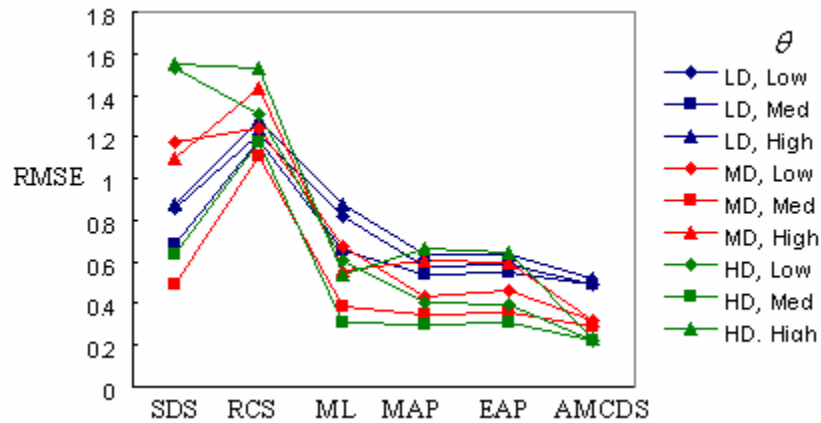
Table 2
Recovery of True Change by the SDS, RCS, IRTDS and AMCDS
as Indexed by r , Combining All Nine Change Conditions

Item Discrimination	SDS	RCS	IRTDS			
			ML	MAP	EAP	AMCDS
LD	.449	.014	.482	.501	.500	.637
MD	.450	.011	.589	.585	.587	.797
HD	.404	.009	.630	.570	.578	.877

Root mean squared error. The RMSE values obtained for recovery of true change by the observed change scores are shown in Figure 7. AMCDS had consistently lowest RMSE of true change with estimated change across all conditions. The results of the repeated-measures ANOVA for the transformed RMSE values (LMSE, Equation 18) are also presented in Table 1. The ANOVA indicated that the largest effect size for the transformed RMSE values was due to

the approach to measuring change, which accounted for 67% of the total variance in the recovery of true change. The mean RMSE values for the SDS, RCS, IRTDS by ML, IRTDS by MAP, IRTDS by EAP, and AMCDS, across the item discrimination test conditions, T1 θ levels, magnitudes of true change, and variabilities of true change, were 0.989, 1.273, 0.602, 0.501, 0.503, and 0.346, respectively. These results indicate that the AMCDS resulted in the lowest RMSE values, and was the approach to measuring change that best recovered true change among all the observed change scores examined, followed by IRTDS, SDS, and RCS.

Figure 7
Recovery of True Change as Indexed by RMSE



Average bias. The average bias values as the index for recovery of true change by the observed change scores are shown in Figure 8. AMCDS was the only method to have essentially zero bias between estimated and true change across all conditions. The results of the repeated-measures ANOVA for the transformed bias values are presented in Table 1. The largest effect size for average bias was due to the approach to measuring change, which accounted for 66% of the total variance in the recovery of true change. The average bias values for the SDS, RCS, IRTDS by ML, IRTDS by MAP, IRTDS by EAP, and AMCDS across all item discrimination test conditions, T1 θ levels, magnitudes of true change, and variabilities of true change, were -0.232 , 1.001 , -0.008 , 0.248 , 0.203 and -0.005 , respectively, indicating that the AMCDS recovered true change among all the observed change scores examined with virtually no bias. Among the observed change scores based on CTs, IRTDS was the approach to measuring change that best recovered true change, followed by SDS, and RCS.

Occurrence of significant change with AMC. In AMC, significant change was defined as non-overlapping SE bands at two measurement occasions. Figure 9 presents the mean number of cases in which significant change was observed and the mean number of items administered before the significant change occurred. As the discrimination of test items increased, the mean number of cases of significant change dramatically increased, from 161 (out of 500) for the LD test condition to 392 cases for the HD test condition (Figure 9a), and the mean number of items administered decreased from 17 items for the LD condition to 11 items for the HD test condition (Figure 9b). Under the HD test condition, when the magnitude of true change was high ($\theta = 1.5$), significant change occurred after administration of an average of only 6 items. The maximum mean number of items administered was 21.34 for the LL change condition with T1 θ level under the LD test condition, which was much less than the 50 items used in the CTs.

Figure 8
Recovery of True Change as Indexed by Bias

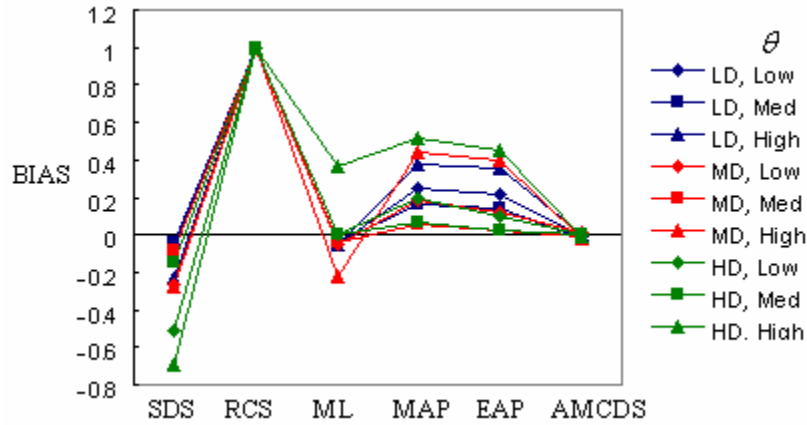
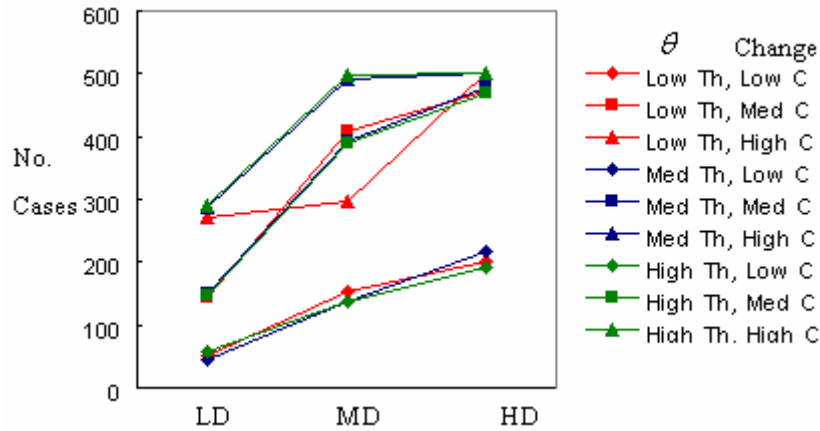
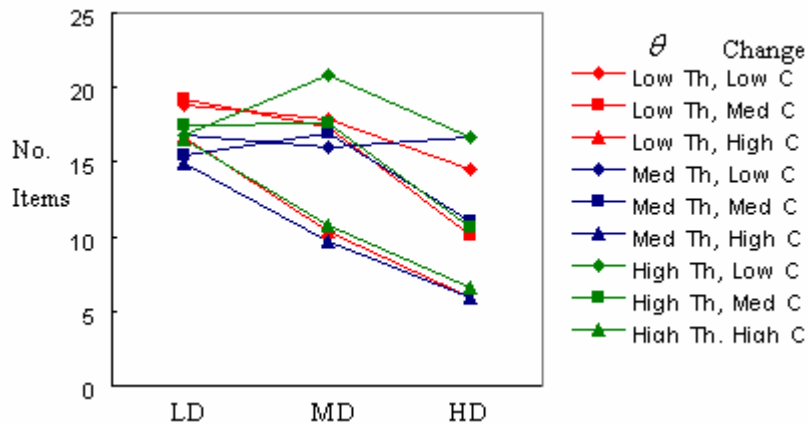


Figure 9
Identification of Significant Change by AMC

a. Mean Number of Cases With Significant Change



b. Mean Number of Items Required to Identify Significant Change



AMC SEs. The mean SE_1 and overall mean SE_2 reflect the mean standard error values after administration of 50 items, while the mean SE_2 for change cases was obtained for the cases in which significant change occurred. As Table 3 shows, the mean SE_1 and overall mean SE_2 were almost identical; however, the SE_2 for significant cases were noticeably larger than the overall SE_2 , as expected. The maximum mean number of items administered for the change cases identified by AMC was only 21.34, which was substantially less than 50 items. Consequently, the resulting SE_2 for significant cases would be larger than the overall SE_2 .

Table 3
SEs for the Estimates of θ_1 and θ_2 by the AMC Tests

Level of θ at Time 1 and Change Condition	LD			MD			HD		
	Mean SE_1	Mean SE_2		Mean SE_1	Mean SE_2		Mean SE_1	Mean SE_2	
		Overall	Change Cases		Overall	Change Cases		Overall	Change Cases
Low θ : LL	0.361	0.353	0.594	0.222	0.215	0.409	0.162	0.154	0.315
LM	0.361	0.352	0.602	0.222	0.215	0.406	0.162	0.154	0.312
LH	0.361	0.353	0.589	0.222	0.214	0.410	0.162	0.154	0.310
ML	0.361	0.346	0.573	0.222	0.209	0.420	0.162	0.150	0.380
MM	0.361	0.347	0.552	0.222	0.210	0.422	0.162	0.151	0.372
MH	0.361	0.347	0.586	0.222	0.210	0.423	0.162	0.151	0.383
HL	0.361	0.345	0.575	0.222	0.208	0.502	0.162	0.149	0.428
HM	0.361	0.345	0.561	0.222	0.208	0.506	0.162	0.150	0.454
HH	0.361	0.347	0.567	0.222	0.208	0.506	0.162	0.150	0.456
Medium θ : LL	0.346	0.349	0.535	0.209	0.206	0.457	0.150	0.149	0.306
LM	0.346	0.348	0.518	0.209	0.207	0.433	0.150	0.149	0.309
LH	0.346	0.348	0.507	0.209	0.206	0.451	0.150	0.149	0.317
ML	0.346	0.352	0.535	0.209	0.208	0.440	0.150	0.150	0.367
MM	0.346	0.353	0.528	0.209	0.208	0.436	0.150	0.150	0.368
MH	0.346	0.352	0.539	0.209	0.208	0.443	0.150	0.151	0.362
HL	0.346	0.359	0.530	0.209	0.214	0.513	0.150	0.155	0.449
HM	0.346	0.358	0.532	0.209	0.213	0.509	0.150	0.155	0.458
HH	0.346	0.358	0.519	0.209	0.214	0.511	0.150	0.156	0.447
High θ : LL	0.359	0.367	0.596	0.213	0.219	0.408	0.156	0.162	0.305
LM	0.359	0.366	0.485	0.213	0.219	0.407	0.156	0.163	0.304
LH	0.359	0.366	0.499	0.213	0.220	0.415	0.156	0.163	0.309
ML	0.359	0.374	0.502	0.213	0.226	0.424	0.156	0.169	0.369
MM	0.359	0.375	0.490	0.213	0.226	0.419	0.156	0.169	0.371
MH	0.359	0.374	0.481	0.213	0.226	0.429	0.156	0.169	0.370
HL	0.359	0.385	0.519	0.213	0.233	0.511	0.156	0.177	0.460
HM	0.359	0.385	0.511	0.213	0.233	0.507	0.156	0.178	0.458
HH	0.359	0.385	0.503	0.213	0.235	0.512	0.156	0.179	0.457

Discussion and Conclusions

The CTs estimated individual change reasonably well when the test was highly discriminating and when the test matched the targeted θ level (i.e., matched the test difficulty) at Time 1, but measured change poorly when the test difficulty and θ were not matched. The standard deviations for the change scores based on CTs were smaller and the correlations between the true and estimated values were higher in the medium θ condition than in either low or high θ conditions. This tendency became more pronounced as the discrimination of the test items increased, indicating that increasing item discrimination on the CTs improved the recovery of true change. However, the results indicate that the conventional tests used to measure change were heavily affected by the range of θ to which the test was targeted and by the item discrimination level of the test.

Among the three conventional approaches to measuring change, the IRTDS generally showed the best recovery of true change, followed by the SDS and the RCS. Among the three IRTDS obtained by the three different estimation methods of estimating θ (ML, MAP and EAP), the ML IRTDS provided the best results for the medium T1 θ group, to which the test was targeted. For the medium T1 θ group, the ML IRTDS reflected the average true change better than the MAP or EAP IRTDS, and the correlations between true θ at T1 and the ML ITRDS were closer to 0.0 than those by MAP and EAP. The correlations of true θ and ML estimated θ at Time 1 and Time 2 were similar to those by MAP and EAP. Beyond the targeted range, the IRTDS was not able to adequately measure change using any of the three θ estimation methods.

None of the approaches to measuring individual change based on the CTs recovered true change better than the AMC. Unlike the CT approaches, which functioned differently at different levels of θ , AMC measured individual change equally well for the middle range of θ as well as for the high and low Time 1 θ . The AMCDS values reflected the average true change values, and the standard deviations for the AMCDS were smaller than any of the CT measures for all levels of θ and for all three test conditions of item discrimination. Similar standard deviations for the AMCDS were found across all θ levels within each item discrimination condition, and standard deviations for the AMCDS became smaller as the item discrimination increased, indicating that higher discriminating items in the AMC produced more precise estimates of individual true change. The standard error values for the θ estimates at Time 1 and 2 were similar across all θ levels within each test condition of item discrimination and had a tendency to decrease as the item discrimination increased (Table 3).

The ANOVA results indicated that approach to measuring change was the only factor, among the main effects and interactions, that had a strong impact on the recovery of change, accounting for 11%, 67%, and 66% of the total variance by r , RMSE, and bias, respectively. The lowest values of RMSE and bias were obtained for the AMCDS, among the approaches examined. Furthermore, the correlations of the true change score and the estimated change score were highest for the AMCDS. These results indicate that estimates of change scores across conditions were closer to the true change scores for the CAT than for any of the conventional tests examined.

Although the adaptive test was designed to have equal length to that of the conventional tests, additional analysis for the occurrence of significant change demonstrated good Time 2 efficiency for AMC. AMC detected significant change—defined by non-overlapping confidence intervals for the two CAT θ estimates—after only 6 to 21 items were administered at Time 2 (versus 50

items for the conventional test). As the test became more discriminating and the mean magnitude of true change was high ($\theta = 1.5$), significant change was detected with an average of only six items.

In addition to providing good Time 2 efficiency, the results support the notion that the AMC provides equal precision for individuals at all levels of the trait being measured across Time 1 and Time 2. The standard error values reported in this study could provide useful information for determining a pre-set SE value as a termination criterion for use in administering adaptive tests to measure change.

The results also provide additional information for designing and implementing adaptive measures of individual change. First, high item discrimination is an important condition when designing an item bank for the AMC, because an item bank with highly discriminating items provided the most precise estimates of individual change. Second, similar standard deviations and measurements of equal precision for simulees were observed for AMC across T1 θ levels within each of the three different item discrimination conditions, indicating that change measurement using AMC would not be affected by Time 1 θ level under each item bank examined. Third, as expected, the magnitude of change is an influential condition in the detection of significant change using AMC. As the magnitude of change became higher, the more cases of significant change occurred within each test condition of item discrimination. Fourth, two different entry levels were compared, with the other CAT components being equal during the administration of the AMC. The use of average θ ($\theta = 0.0$) at Time 1 as the starting value and the variable entry at Time 2 (using the final ML θ estimate at Time 1), reported similar results in terms of their correlations between the true and observed AMC score (Figure 3) and their standard error values (Table 3). The use of average θ as a starting value at Time 1 would be a reasonable choice unless prior information is available. However, the use of the maximum likelihood estimate of θ obtained at Time 1 as the entry level at Time 2 played a role to make CAT more efficient by reducing the number of items administered in AMC under the high discrimination test condition. Finally, the maximum number of items (50 items) criterion was used as the termination criterion during the administration of AMC at both measurement occasions to enable direct comparisons of results from CTs and the AMC. However, using non-overlapping SE bands as a termination criterion to detect significant change, the average number of items administered was at most 21, which was much less than 50 items. This result supported the use of non-overlapping SE bands as a termination criterion in measuring individual change, but additional research on other termination criteria for the AMC is needed.

The results of this study indicate that AMC is a viable and effective method for measuring individual change. It performed best for all criteria examined in this study. In addition, AMC is efficient—it can dramatically reduce the number of items necessary to measure individual change. AMC demonstrated substantial advantages over conventional testing approaches for measuring individual change, in terms of the recovery of true change under the conditions examined. The conditions delineated in this study were restricted to measuring individual change at two points in time. In addition, the item banks and testing procedures were designed to fairly evaluate the performance of conventional tests for measuring change, thus the banks were not necessarily ideal for an adaptive testing environment (e.g., the peaked-rectangular item banks) and termination at 50 items might be suboptimal for CATs.

More extensive research on AMC is indicated. For example, what are the optimal conditions

for measuring individual change with AMC when examining change over more than two points in time? What kind of item bank is required to accurately measure change at multiple occasions? If significant change has not occurred between measurement occasions, what additional termination criteria might be appropriate for multiple occasions of measurement? Since infinitely large item banks are not possible, especially in a classroom application, results from future research could provide useful information for making instructional decisions more appropriate at the individual level and might lead to a coherent integration between teaching and testing.

References

- Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.
- Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika*, *50*, 3–16.
- Assessment Systems Corporation. (1997). *User's manual for the PARDSIM parameter and response data simulation program*. St. Paul, MN: Assessment Systems Corporation.
- Assessment Systems Corporation. (1998). *User's manual for the conventional test scoring program*. St. Paul, MN: Assessment Systems Corporation.
- Assessment Systems Corporation. (2005). *Manual for POSTSIM 2.0: Post-hoc (real-data) simulation of computerized adaptive testing*. St. Paul, MN: Assessment Systems Corporation.
- 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.
- Cronbach, L. J., & Furby, L. (1970). How we should measure “change”—or should we? *Psychological Bulletin*, *74*, 68–80.
- Duncan, A. D. (1974). Tracking behavioral growth: Day-to-day measure of frequency over domain of performance. *Educational Technology*, *14*, 54–59.
- 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.
- 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.
- Friedenberg, L. (1995). *Psychological testing: Design, analysis, and use*. Boston: Allyn and

Bacon.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.

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

Hays, W. L. (1988). *Statistics* (4th ed.). New York: Holt, Rinehart & Winston.

Howell, D. C. (1992). *Statistical methods for psychology* (3rd ed.). Boston: PWS-KENT Publishing company.

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).

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* (pp. 257–283). New York: Academic Press.

Lord, F. M. (1958). The utilization of unreliable change scores. *Journal of Educational Psychology*, 3, 150–152.

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., & 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.

Tinkelman, S. N. (1971). Planning the objective test. In R.L. Thorndike (Ed.), *Educational Measurement*, (2nd ed.) (pp. 46–80). Washington, D.C.: American Council on Education.

Traub, R. E. (1967). A note on the reliability of residual change scores. *Journal of Educational Measurement*, 4, 253–256.

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

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

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, 6, 473–492.

Weiss, D. J. (1983). Introduction. In D.J. Weiss (Ed.), *New horizons in testing* (pp. 1–8). New York: Academic Press

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical*

Psychology, 53, 774–789.

Weiss, D. J. (1995). Improving individual differences measurement with item response theory and computerized adaptive testing. In D. Lubinski & R. V. Dawis (Eds.), *Assessing individual differences in human behavior: New concepts, methods, and findings* (pp. 49-79). Palo Alto CA: Davies-Black Publishing.

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Evaluation in Counseling and Development*, 37, 70–84.

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.

Yoes, M. E. (1993). *A comparison of the effectiveness of item parameter estimation techniques used with the 3-parameter logistic item response theory model*. Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

Zwick, R., Thayer, D. T., & Wingersky, M. (1994). A simulation study of methods for assessing differential item functioning in computerized adaptive testing. *Applied Psychological Measurement*, 18, 121–140.