Adaptive Measurement of Individual Change

Gyenam Kim-Kang[1] and David J. Weiss[2]

[1]Korea Nazarene University, Cheonan, ChungNam, South Korea
[2]University of Minnesota, Minneapolis, MN, USA

Abstract. Adaptive measurement of change (AMC) was investigated by examining the recovery of true change. Monte Carlo simulation was used to compare three conventional testing (CT) methods with AMC. The CTs estimated individual change moderately well when the test was highly discriminating and when the θ level matched the test difficulty. However, AMC measured individual change equally well across the entire range of θ. AMC with more discriminating items produced the most precise estimates of individual change. AMC was shown to be superior to CTs under all conditions examined. In addition, AMC is efficient – it can dramatically reduce the number of items necessary to measure individual change. The results indicate that AMC is a viable and effective method for measuring individual change.

Keywords: computerized adaptive testing, measuring change, item response theory, residual change score, difference score

Introduction

The measurement of individual change has been one of the fundamental concepts in psychological and educational research. Education is primarily intended to produce learning, which should result in changes in achievement level for each student. In measuring individual change a student's progress can be ascertained by comparing that student's previous status with their current status on some achievement variable. Similarly, indicators of change are also important in other areas of psychology. For example, in measuring an individual's response to some psychological clinical treatment program, it is important to know whether the individual's level on a relevant variable (e.g., depression) has increased, decreased, or remained constant. However, the measurement of change at the individual level has been one of the most controversial issues over the years (e.g., Bereiter, 1963; Cronbach & Furby, 1970; Embretson, 1995). One of the traditional ways to measure individual change is simply to compute the difference between measurements obtained at two points in time, such as the pretest-posttest paradigm. This simple difference score is given by (Burr & Nesselroade, 1990; McDonald, 1999)

$$D_j = Y_j - X_j \tag{1}$$

where D_j is the observed change or difference score for person j, Y_j is the observed score at Time 2, and X_j is the observed score at Time 1.

Previous research regarding the simple difference score for the measurement of individual change has identified major problems:

1) low reliability (Allen & Yen, 1979; Embretson, 1995; Hummel-Rossi & Weinberg, 1975; Lord, 1963; Willett, 1994, 1997),
2) negative correlation between change scores and initial status (Cronbach & Furby, 1970; Embretson, 1995; Willett, 1994, 1997),
3) regression toward the mean (Cronbach & Furby, 1970; Hummel-Rossi & Weinberg, 1975), and
4) dependence on the scale of measurement employed at two or more points of measurement (Embretson, 1995; Hummel-Rossi & Weinberg, 1975).

Several different procedures for estimating change have been suggested (Lord, 1963; Manning & DuBois, 1962; 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, 1997). Manning and DuBois showed theoretically that the RCS is more reliable than the simple difference score in most situations. The RCS reflects the difference between an actual and a predicted score and is given as

$$R_j = Y_j - Y_j' \tag{2}$$

$$R_j = Y_j - \overline{Y} - b_{Y.X}\,(X_j - \overline{X}) \tag{3}$$

where, Y_j is the observed score at Time 2 for person j, X_j is their observed score at Time 1, Y_j' is the predicted score from X_j based on the bivariate linear regression of Y on X, $\overline{X}$ and $\overline{Y}$ 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.

DOI 10.1027/0044-3409.216.1.49

In order to obtain the RCS, group level information is required to estimate the regression of *Y* on *X*. 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 predicted value. 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) and has the potential to reduce several of the problems inherent in using conventional tests to measure individual change. A number of researchers have addressed the issue of measuring change using item response theory (IRT) models. Fischer (1976) proposed the linear logistic latent trait model, Bock (1976) developed an IRT model for growth curves, Andersen (1985) developed a multidimensional Rasch model for repeated testings, and Embretson (1991a, 1991b) proposed a multidimensional Rasch model for learning and measuring change. However, 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. 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 and IRT 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.

Weiss and Kingsbury (1984) proposed a method for measuring individual change (which they referred to as adaptive self-referenced testing) that combined the benefits of both IRT and computerized adaptive testing (CAT). The characteristics of CATs are that different items, or sets of items, are administered to different individuals depending on each individual's status on the latent trait as it is continuously estimated during test administration (Weiss, 1982, 1983, 1985, 1995, 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).

Weiss and Kingsbury's method, referred to here as adaptive measurement of change or 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. In AMC, the measurement of change for a particular examinee is determined with reference to that examinee's previous trait level estimate. Using AMC, change is measured as the difference between an individual's estimated levels of the trait θ_j ($\hat{\theta}_j$) for two (or more) occasions. Significant change is said to occur when the IRT-based confidence intervals around the $\hat{\theta}_j$s for two estimates do not overlap (Weiss & Kingsbury, 1984). The confidence intervals, or standard error (*SE*) bands, are generally approximated as

$$\theta_j \pm 2\left(SE \mid \hat{\theta}_j\right), \tag{4}$$

where $SE \mid \hat{\theta}_j$ is determined from the second derivative of the log-likelihood function (Weiss, 2005, pp. 10–11; Baker, 1992, pp. 69–72),

$$SE \mid \hat{\theta}_j = \sqrt{Var(\hat{\theta}_j \mid \theta_j)}\,, \tag{5}$$

where, asymptotically

$$Var(\hat{\theta}_j \mid \theta_j) \approx \frac{1}{I(\hat{\theta}_j)} \tag{6}$$

and $I(\hat{\theta}_j)$ is obtained by substituting $\hat{\theta}$ for θ after the second derivative is taken in

$$I(\theta_j) = \left(\frac{\partial^2 \ln L_j}{\partial \theta_j^2} \right). \tag{7}$$

The measurement of change for an individual by AMC is determined with reference only to that individual, reflecting how the individual's $\hat{\theta}$ at Time 2 differs from their own $\hat{\theta}$ 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 conventional test 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. In addition, the study was a first examination of the power of AMC to detect different magnitudes of 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 the AMC IRT-based CAT approach. 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 Score Distribution

The T1 true θ values of 1,500 simulated examinees were generated to have a rectangular distribution with mean 0.0, 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 (500 simulees in each group) – low (–2.25 to –0.75), me-

dium (–0.7499 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 scores. For each of the three initial 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). The true θ value at T2 was 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

$$P_{ij}\left(\theta_j\right) = c_i + \left(1 - c_i\right)\frac{\exp[a_i(\theta_j - b_i)]}{1 + \exp[a_i(\theta_j - b_i)]}, \qquad (8)$$

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

The three different average item discrimination conditions were low (LD; $\overline{a}$ = 0.5), medium (MD; $\overline{a}$ = 1.0), and high (HD; $\overline{a}$ = 1.5), respectively, with average standard deviation (SD) of 0.15 in each item bank. The distribution of item difficulties was centered at approximately b_i = 0.00 and had 18 intervals from b_i = –4.50 to b_i = +4.50, with a range wider than the true T1 θ range [θ = –2.25 to +2.25]. The middle six intervals, θ = –1.5 to +1.5, contained 24 items, while the other 12 intervals (the lowest and highest six intervals) contained only 12 items per interval. Therefore, each item bank consisted of 288 items, with more items available in the middle range for conventional tests drawn from the same item bank. The pseudo-guessing parameter for all the items was fixed at 0.20 for all test conditions (Kingsbury & Weiss, 1983; Lord & Novick, 1968; Urry, 1977; Yen, 1986).

Conventional Test

The conventional tests (CTs) were constructed as parallel tests to measure individual change. From each item bank, two parallel 50-item fixed-length CTs were constructed with item difficulties ranging from –1.5 to +1.5 (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 of b_i, containing 24 items in each interval in the corresponding item bank, ranging from b_i = –1.5 to b_i = +1.5 with mean b_i = 0.0. Items for the parallel test were selected at random from those items not previously selected in each item bank. A total of six CTs (three sets of parallel forms) 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 of the test for each simulee was generated using Equation 8. 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 between 0 and 1. A correct answer (1) was recorded for a simulee if the random number was less than the model-generated probability; otherwise, the item was scored as incorrect (0).

The number-correct (NC) scores and the maximum likelihood (ML) estimates of θ ($\hat{\theta}_{C1}$ and $\hat{\theta}_{C2}$) 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, 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 parameters for all 288 items using PARDSIM (Assessment Systems Corporation, 1997). The procedure to generate the item responses was the same as for the CTs, except that the number of items was different (288 versus 50). As a result, a 1,500 × 288 item response matrix was constructed from which CATs were administered.

The administration of CAT was performed using POSTSIM, a program for posthoc simulation of CAT (Weiss, 2005). In administering CAT, the initial θ value was the same (θ = 0) for all simulees at T1. ML estimation (Baker, 1992; Yoes, 1993; Zwick, Thayer, & Wingersky, 1994) was used to estimate $\hat{\theta}$ at both T1 ($\hat{\theta}_{A1}$) and T2 ($\hat{\theta}_{A2}$). However, ML estimation cannot be used for nonmixed response patterns (all correct response[s] or all incorrect response[s]), so a step size of ±3 was used to select the next item until a mixed response pattern (at least one correct and one incorrect response) was obtained. 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). The CAT procedure was terminated after the administration of 50 items at both T1 and T2 to enable

a direct comparison by matching the number of items in the CTs. The final ML estimation of θ obtained at T1 ($\hat{\theta}_{A1}$) was used as the entry level for the T2 CAT.

Identifying Significant Change with AMC

Based on the final MLE of θ at T1 ($\hat{\theta}_{A1}$) and the results of administration of CATs after each item at T2, the power of AMC was evaluated by determining the number/proportion of cases in which significant change was observed (defined as nonoverlapping SE bands at two measurement occasions), for each change condition in each T1 θ level for each of three item discrimination conditions. In addition, the mean number of items administered at T2 and the mean SE values ($SE_{2,\mathrm{AMC}}$) for only those cases with significant change were determined.

Approaches to Measuring Change

Four approaches to measuring change were examined. For the CTs, three scores were used: (1) the simple difference score (SDS) was obtained by substituting the transformed value of NC scores to the θ metric at T2 ($\hat{\theta}_{NC2,j}$) and at T1 ($\hat{\theta}_{NC1,j}$) into Equation 1; (2) the RCS was similarly operationalized based on Equations 2 and 3, using transformed values of NC scores to the θ metric, $\hat{\theta}_{NC2,j}$ and $\hat{\theta}_{NC1,j}$, corresponding to the NC scores at T2 and T1, respectively, and

$$b\left(\hat{\theta}_{NC2,j}\cdot\hat{\theta}_{NC1,j}\right)=r\left(\hat{\theta}_{NC2,j}\hat{\theta}_{NC1,j}\right)\left[\frac{s\left(\hat{\theta}_{NC2,j}\right)}{s\left(\hat{\theta}_{NC1,j}\right)}\right]; \quad (9)$$

(3) the IRT-scored difference score (IRTDS) was defined as the difference between $\hat{\theta}_{C1,j}$ and $\hat{\theta}_{C2,j}$ using ML $\hat{\theta}$s. For AMC, the difference score (AMCDS) was defined as the difference between the ML $\hat{\theta}$s at T1 and T2 ($\hat{\theta}_{A1,j}$ and $\hat{\theta}_{A2,j}$).

Evaluation Criteria

Recovery of True Change

How well each of the four approaches recovered true change was evaluated using Pearson product-moment correlation coefficients, root mean square error (RMSE), and the average bias (BIAS) between true and estimated change values for each of the 27 different change conditions across initial (T1) θ level, within each of the three different item discrimination conditions. In computing these indices, d_j was the true change value ($\theta_{2j}-\theta_{1j}$) and $\hat{d}_j$ was each of the observed change values (SDS, RCS, IRTDS, and AMCDS). Thus, the correlation was computed as $r(d_j,\hat{d}_j)$, RMSE was computed as

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{j=1}^{N}\left(d_j-\hat{d}_j\right)^2}{N}}, \quad (10)$$

and BIAS was computed as

$$\mathrm{BIAS}=\frac{\sum_{j=1}^{N}\left(d_j-\hat{d}_j\right)}{N}. \quad (11)$$

Positive BIAS indicates underestimating the true change and a negative value reflects overestimating true change.

Effect Sizes

Effect sizes were computed for each of the three evaluative criteria. The study design was a repeated measures ANOVA for each of the three evaluative criteria (Howell, 1992) 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.

Effect size was calculated as

$$\eta^2=\frac{SS_{Effect}}{SS_{total}}, \quad (12)$$

where SS_{Effect} is the sum of squares 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):

$$r_Z=\frac{1}{2}\cdot\ln\left(\frac{1+r}{1-r}\right), \quad (13)$$

$$\mathrm{LMSE}=\log_{10}(RMSE+1),\ \text{and} \quad (14)$$

$$\mathrm{LBIAS}=\log_{10}(BIAS+1) \quad (15)$$

Results

Descriptive Statistics

Table 1 shows that the mean values of estimated θ from the CTs ($\hat{\theta}_{NC1}$, $\hat{\theta}_{NC2}$, $\hat{\theta}_{C1}$, $\hat{\theta}_{C2}$) closely approximated the corresponding true T1 θ and T2 θ values, respectively, and their *SD*s were close to those of true T1 and T2 θ values for the medium T1 θ level. However, for the low and high θ group, mean CT $\hat{\theta}$s deviated from the corresponding true mean values, and their *SD*s were larger. This trend was pronounced for the MD and HD test condition. The mean ML $\hat{\theta}_{C1}$ and $\hat{\theta}_{C2}$ values were substantially smaller than true θs for the high T1 θ level of the HD test condition–1.38 for $\hat{\theta}_{C1}$ and 1.82 for $\hat{\theta}_{C2}$.

However, Table 1 also shows that for AMC the observed mean $\hat{\theta}_{A1}$ and $\hat{\theta}_{A2}$ reflected the corresponding true θ values (θ_1,θ_2) across all conditions. The *SD*s of $\hat{\theta}_{A1}$ and $\hat{\theta}_{A2}$ were also very similar across all conditions.

Table 1. Average of means and *SD*s of true and estimated θs across low, medium, and high variability of change conditions for combinations of LD, MD, and HD tests and low, medium, and high θ groups

Test condition and θ_1 level		θ_1		θ_2		$\hat{\theta}_{NC1}$		$\hat{\theta}_{NC2}$		$\hat{\theta}_{C1}$		$\hat{\theta}_{C2}$		$\hat{\theta}_{A1}$		$\hat{\theta}_{A2}$	
		Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
LD	Low θ	–1.501	0.424	–0.500	0.428	–1.583	0.840	–0.527	0.690	–1.564	0.807	–0.518	0.675	–1.504	0.542	–0.507	0.555
	Med θ	0.001	0.429	1.001	0.434	0.005	0.652	1.045	0.678	0.010	0.633	1.048	0.662	0.004	0.549	1.021	0.561
	High θ	1.503	0.446	2.504	0.450	1.588	0.713	2.620	0.820	1.578	0.693	2.621	0.847	1.531	0.555	2.529	0.594
MD	Low θ	–1.501	0.424	–0.500	0.428	–1.844	1.213	–0.574	0.683	–1.532	0.725	–0.514	0.547	–1.493	0.472	–0.502	0.483
	Med θ	0.001	0.429	1.001	0.434	–0.014	0.501	1.072	0.629	–0.012	0.497	1.026	0.532	0.006	0.493	1.011	0.489
	High θ	1.503	0.446	2.504	0.450	1.628	0.791	2.901	1.064	1.511	0.531	2.162	0.395	1.500	0.500	2.516	0.505
HD	Low θ	–1.501	0.424	–0.500	0.428	–2.110	1.485	–0.603	0.752	–1.396	0.700	–0.488	0.487	–1.497	0.462	–0.495	0.457
	Med θ	0.001	0.429	1.001	0.434	–0.003	0.505	1.150	0.808	0.009	0.488	0.971	0.464	0.000	0.456	1.005	0.460
	High θ	1.503	0.446	2.504	0.450	1.835	1.168	3.529	1.135	1.378	0.403	1.817	0.205	1.509	0.473	2.514	0.482

Table 2. Average correlations of observed Time 1 and Time 2 scores, and of true and observed Time 1 and Time 2 scores, across low, medium, and high variability of change conditions for combinations of LD, MD, and HD tests and low, medium, and high θ groups

Test condition & θ_1 level		$(\hat{\theta}_{NC1}, \hat{\theta}_{NC2})$	$(\hat{\theta}_{C1}, \hat{\theta}_{C2})$	$(\hat{\theta}_{A1}, \hat{\theta}_{A2})$	$(\theta_1, \hat{\theta}_{NC1})$	$(\theta_1, \hat{\theta}_{C1})$	$(\theta_1, \hat{\theta}_{A1})$	$(\theta_2, \hat{\theta}_{NC2})$	$(\theta_2, \hat{\theta}_{C2})$	$(\theta_2, \hat{\theta}_{A2})$
LD	Low θ	0.390	0.402	0.590	0.575	0.587	0.763	0.667	0.683	0.768
	Med θ	0.465	0.485	0.598	0.698	0.712	0.776	0.675	0.692	0.763
	High θ	0.353	0.372	0.586	0.627	0.656	0.775	0.543	0.542	0.762
MD	Low θ	0.407	0.468	0.771	0.536	0.579	0.873	0.757	0.814	0.893
	Med θ	0.663	0.713	0.791	0.859	0.873	0.899	0.784	0.828	0.896
	High θ	0.370	0.384	0.790	0.668	0.780	0.901	0.548	0.517	0.894
HD	Lowθ	0.368	0.455	0.873	0.486	0.530	0.939	0.752	0.872	0.941
	Med θ	0.672	0.775	0.871	0.898	0.913	0.936	0.760	0.861	0.942
	High θ	0.281	0.260	0.873	0.646	0.793	0.946	0.461	0.358	0.935

Table 3. Average of means and SDs of change scores across low, medium, and high variability of change conditions for combinations of LD, MD, and HD tests and low, medium, and high θ groups

Test condition and θ_1 level		True change		SDS		RCS		IRTDS		AMCDS	
		Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
LD	Low θ	1.001	0.054	1.056	0.856	0.000	0.635	1.050	0.818	1.010	0.774
	Med θ	1.001	0.054	1.040	0.689	0.000	0.600	1.039	0.659	1.017	0.498
	High θ	1.001	0.054	1.022	0.877	0.000	0.767	1.050	0.873	0.998	0.525
MD	Low θ	1.001	0.054	1.271	1.235	0.000	0.626	1.047	0.635	0.994	0.324
	Med θ	1.001	0.054	1.085	0.484	0.000	0.474	1.045	0.389	1.005	0.315
	High θ	1.001	0.054	1.273	1.065	0.000	0.986	0.772	0.498	1.016	0.326
HD	Low θ	1.001	0.054	1.507	1.438	0.000	0.704	0.995	0.634	1.003	0.232
	Med θ	1.001	0.054	1.152	0.624	0.000	0.612	0.995	0.315	1.005	0.233
	High θ	1.001	0.054	1.694	1.381	0.000	1.075	0.625	0.360	1.005	0.241

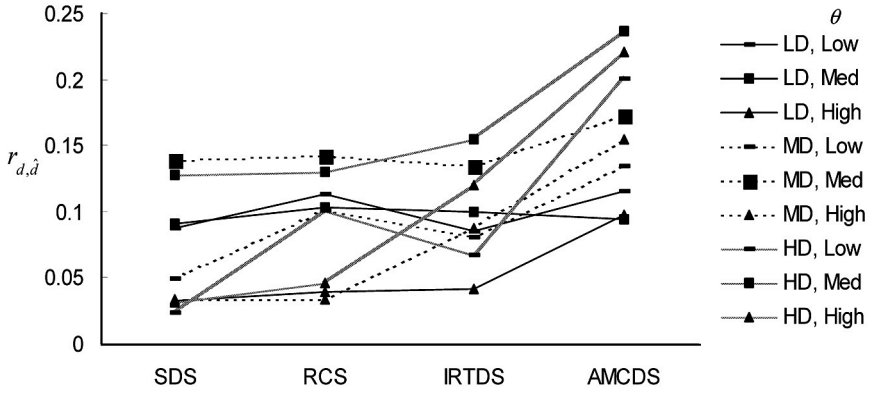

Figure 1. Recovery of true change as indexed by *r*.

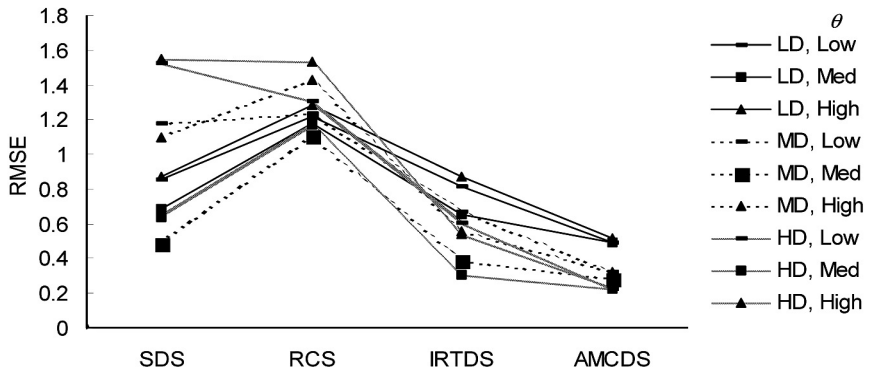

Figure 2. Recovery of true change as indexed by RMSE.

Correlations Between Observed T1 and T2 Scores, and Between True and Observed Scores

Table 2 shows that the correlations between two observed scores and between true θs and observed scores at T1 and T2, based on CTs, were higher for the medium T1 θ group (.47 to .91) than for the low (.37 to .87) and high T1 θ group (.28 to .79) across all item discrimination test conditions. The correlations involving the estimates from adaptive testing ($\hat{\theta}_{A1}$ & $\hat{\theta}_{A2}$, $\hat{\theta}_1$ & $\hat{\theta}_{A1}$, and $\hat{\theta}_2$ & $\hat{\theta}_{A2}$), also shown in Table 2, were higher than those from CTs for all conditions and increased as the discrimination of test items increased. Furthermore, unlike the correlation results from CTs, those from adaptive tests were similar across the true T1 θ levels.

Change Scores

The mean AMCDS values reflected the average true change values across all conditions (Table 3). Furthermore, the *SD*s of AMCDS were much smaller (.23 to .77) than those from CTs (SDS, RCS, IRTDS) for most T1 θ levels and for all item discrimination test conditions (.32 to 1.44). The *SD*s of AMCDS decreased as the discrimination of test items increased; however, they were still noticeably larger than those of true change (.05).

Table 3 also shows that the mean change scores of SDS and IRTDS reflected the average true change values for the medium T1 θ level and for the low discrimination (LD) test condition (Table 3). The *SD*s for the CTs (0.31 to 1.44) were uniformly higher than those of AMCDS (.23 to .77) and those of the average true change (0.05). The mean RCS values were zero for all conditions. In comparing the observed change scores, these results showed that AMC best reflected the average true change values and resulted in the smallest *SD*s across all conditions examined in this study.

Recovery of True Change

Correlations

Figure 1 shows results for Pearson correlations as the index for recovery of true change by the observed change scores for each of the nine test conditions. AMCDS had consistently highest correlations of true change with estimated

Table 4. Recovery of true change by the SDS, RCS, IRTDS, and AMCDS, combining the three θ levels and the nine change conditions, as indexed by the product-moment correlation (*r*)

Item discrimination	SDS	RCS	IRTDS	AMCDS
LD	0.449	0.014	0.482	0.637
MD	0.450	0.011	0.589	0.797
HD	0.474	0.009	0.630	0.877

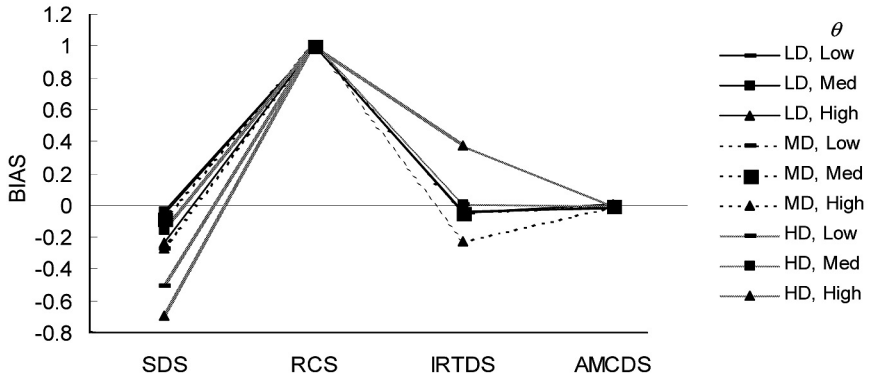

Figure 3. Recovery of true change as indexed by BIAS.

change across all conditions. However, the correlations for all methods, including AMCDS, were quite low, reaching a maximum of .24, because of restriction of range. Table 4 shows Pearson correlations combining all θ the levels and the change conditions for each item discrimination condition. As Table 4 shows, correlations of estimated change with true change were considerably higher than those in Figure 1, ranging from 0.449 for the LD condition for SDS to .877 for AMCDS for the HD condition. Table 4 also shows that IRTDS improved the recovery of true change relative to SDS, and RCS did not recover true change at all.

The results of the repeated-measures ANOVA for the transformed *r* values indicated that the largest effect size was due to the variability of true change, which accounted for 45% of the total variance in the recovery of true change. Mean *r* for low, medium, and high levels of variability of true change, across all item discrimination test conditions, T1 θ levels, magnitudes of true change, and 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, as indexed by *r*.

The effect having the second largest effect size for the transformed *r* was 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, and AMCDS across all other conditions were .068, .090, .097, and .159, respectively, indicating that AMCDS best recovered true change among all the observed change scores examined, followed by IRTDS, RCS, and SDS; the low values of these correlations were the result of restriction of range within each θ group.

Root Mean Squared Error

RMSE values obtained for recovery of true change by the observed change scores are shown in Figure 2. 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, and AMCDS, across the item discrimination test conditions, T1 θ levels, magnitudes of true change, variabilities of true change, were 0.989, 1.273, 0.602, and 0.346, respectively, indicating that the AMCDS resulted in the lowest RMSE value (.346) and, therefore, was the approach to measuring change that best recovered true change among all the change scores examined, followed by IRTDS, SDS, and RCS.

Average BIAS

The average BIAS values as the index for recovery of true change by the observed change scores are presented in Figure 3. AMCDS was the only method to have essentially zero BIAS between estimated and true change across all conditions. The largest effect size for the transformed BIAS values 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, and AMCDS across all item discrimination test conditions, T1 θ levels, magnitudes of true change, variabilities of true change, were –0.232, 1.001, –0.008, and –0.005, respectively, indicating that the AMCDS recovered true change with virtually no bias. Among the observed change scores based on CTs, IRTDS was the approach to measuring individual change that best recovered true change, followed by SDS and RCS.

Identifying Significant Change With AMC

In AMC, significant change was defined as nonoverlapping *SE* bands at two measurement occasions. Figure 4a presents the mean number of cases in which significant changes were observed, and Figure 4b shows the mean number of items administered to identify significant change, for three θ groups and three levels of true change. As item discrimination increased, the mean number of cases of significant change (out of 500) substantially increased, from 161 cases for the LD test condition (power = 0.32) to 392 cases (power = 0.784) for the HD test condition (Figure 4a). At the same time, the mean number of items administered decreased from 17 items for the LD to 11 items for

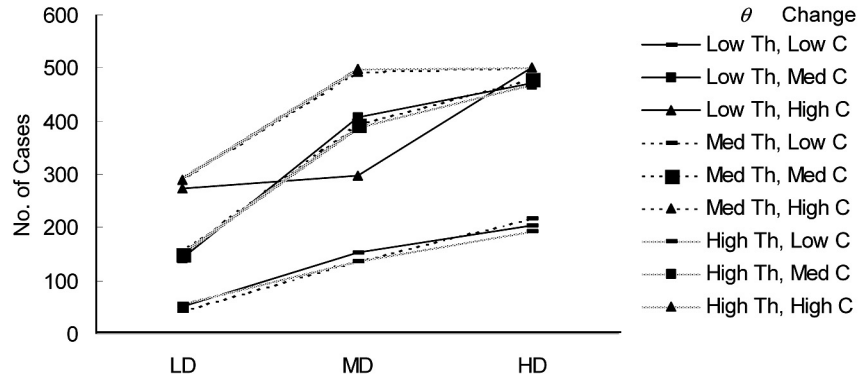

Figure 4a. Mean number of cases of significant change identified by AMC.

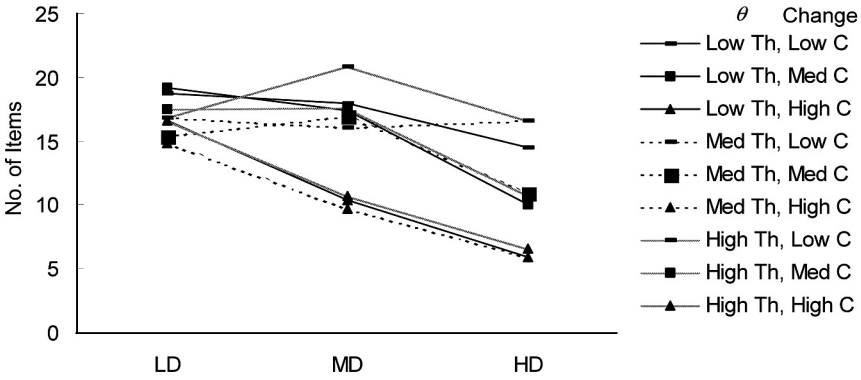

Figure 4b. Mean number of items required at T2 for AMC to identify significant change (50 items were administered at T1).

the HD test condition (Figure 4b). Under the HD test condition, when the magnitude of true change was high (θ = 1.5), significant change was identified after administration of an average of only about 6 items (Figure 4b) with power approximating 1.0.

The mean $SE(\theta_{2,50})$ was computed for the full-length (50-item) T2 CATs and compared to $SE_{2,AMC}$ for the simulees for which significant change was identified by AMC. Mean $SE(\theta_{2,50})$ ranged from .149 to .385. Mean $SE_{2,AMC}$ were somewhat larger, ranging from .305 to .602, with substantially fewer items (an average of 6 to 21 items vs. 50 items).

Discussion and Conclusions

The CTs estimated individual change reasonably well when the tests were highly discriminating and when the θ level matched the test difficulty at T1. The *SD*s 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 the low or high θ condition. The CTs recovered true change best for the medium θ level at T1 and 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. These results suggest that CTs measure individual change best when the range of θ is targeted to the item difficulty level of a test. However, none of the approaches to measuring individual change based on the CTs recovered true change better than AMC.

Unlike the CT approaches, which functioned differently at different levels of θ, AMC measured individual change equally well for all T1 θ level groups. AMCDSs reflected the average true change values, and the *SD*s 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 *SD*s for the AMCDS were observed across all θ levels within each item discrimination condition, and they decreased as item discrimination increased, indicating that more discriminating items in the AMC produced more precise estimates of change at the individual level.

The ANOVA results indicated that approaches 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. AMCDS correlations were the highest among methods for measuring change. These results indicate that estimates of change scores across conditions were closer to the true change scores for the adaptive measures than for any of the measures of change based on conventional tests.

Analysis of the occurrence of significant change indicated good T2 efficiency in that AMC detected significant change – defined by nonoverlapping confidence intervals

for the two CAT θ estimates – after only 6 to 21 items were administered at T2 (vs. 50 items for the CTs). As the test became more discriminating and the mean magnitude of true change was high (1.5 in θ units), significant change was detected with an average of only six items administered. Mean AMC SEs for the significant change cases were similar to those based on 50 items, even with the substantially reduced test lengths. The power of AMC to detect true change was highest with highly discriminating items; in this condition for high and medium change across the entire range of θ, power to detect true change was between 0.94 and 1.00, with an average of fewer than 12 items.

The results of this study have implications for the design and implementation of AMC. First, high item discrimination, accompanied by wider range of item difficulty (b_i = –4.5 to +4.5) than the true θ range (θ = –2.25 to +2.25), is an important condition when designing an item bank for AMC, since an item bank with highly discriminating items provided the most precise estimates of individual change and high power in detecting true change under several combinations of conditions. Second, similar *SD*s and measurements of equal precision for simulees for AMC were reported across T1 θ levels within each of the three different item discrimination conditions, indicating that change measurement using AMC would not be affected by T1 θ level. Third, as expected, the magnitude of change is an influential condition for the detection of significant change by AMC. As the magnitude of true change increased, AMC was able to detect more cases of significant change and the significant changes could be detected with fewer items. 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 a direct comparison of results from CTs and from AMC. However, when nonoverlapping SE bands were used as an AMC termination criterion, the average number of items administered was at most 21– much less than 50 items. This result supported the use of nonoverlapping SE bands as a termination criterion in measuring individual change, but more refined research for appropriate termination criteria for the AMC is required to increase the power of AMC to detect true change, particularly with low discriminating items.

Conclusions

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 was shown to be superior to conventional testing approaches for measuring individual change in terms of the recovery of true change under the conditions examined. The conditions of this study were restricted to measuring individual change at two points in time using an item bank that favored a conventional testing strategy, in addition to fixed CAT termination at 50 items.

More extensive research on AMC needs to be performed, including identification of the optimal conditions for measuring individual change when examining change over more than two points in time. This would include determining the characteristics of the item bank required to accurately measure individual change at multiple occasions. It is also important to determine what additional termination criteria might be appropriate for multiple occasions of measurement if significant change has not occurred between measurement occasions; for example, terminating a CAT when θ estimates and/or their standard errors have stabilized.

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: Author.

Assessment Systems Corporation. (1998). *User's manual for the conventional test scoring program*. St. Paul, MN: Author.

Baker, F.B. (1992). *Item response theory: Parameter estimation techniques.* New York: Marcel Dekker.

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 & L.J. T. van der Kamp (Eds.), *Advances in psychological and educational measurement* (pp. 75–96). New York: Wiley.

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.

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, DC: 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: Wiley.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response the-*

ory: Principles and applications. Boston, MA: Kluwer-Nijhoff.

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

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

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: Latent trait test theory and computerized adaptive testing* (pp. 257–283). New York: Academic Press.

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.

Tinkelman, S.N. (1971). Planning the objective test. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 46–80). Washington, DC: 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: Latent trait test theory and computerized adaptive 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.

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. (2005). *Manual for POSTSIM: Posthoc simulation of computerized adaptive testing. Version 2.0.* St. Paul, MN: Assessment Systems Corporation.

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.

Gyenam Kim Kang

Center for Teaching and Learning
Korea Nazarene University
456 Ssangyong-dong
Cheonan, ChungNam
South Korea 330-718
Tel. +1 765 490 7870
E-mail gnkang35@yahoo.com

