Journal of Computerized Adaptive Testing

Volume 10 Number 3

July 2023

How Do Trait Change Patterns Affect the Performance of Adaptive Measurement of Change?

Ming Him Tai, Allison W. Cooperman, Joseph N. DeWeese, and David J. Weiss

DOI 10.7333/2307-1003032

The *Journal of Computerized Adaptive Testing* is published by the International Association for Computerized Adaptive Testing

ISSN: 2165-6592

International Association for Computerized Adaptive Testing
IACAT
Advancing the Science and Practice of Human Assessment

Journal of Computerized Adaptive Testing

Volume 10, Number 3, July 2023

DOI 10.7333/2307-1003032

ISSN 2165-6592

How Do Trait Change Patterns Affect the Performance of Adaptive Measurement of Change?

Ming Him Tai, Allison W. Cooperman, Joseph N. DeWeese, and David J. Weiss
University of Minnesota

Adaptive measurement of change (AMC) is a psychometric procedure to detect intra-individual change in trait levels across multiple testing occasions. However, in studying how AMC performs as a function of change, most previous studies did not specify change patterns systematically. Inspired by Cronbach and Gleser (1953), a quantitative framework was proposed that systematically decomposes a change pattern into three components: magnitude, scatter, and shape. Shape was further decomposed into direction and order. Using monte-carlo simulations, a series of analyses of variance were performed to investigate how each of these components affected the false positive rates (FPRs) and true positive rates (TPRs) for detecting true change, and a change recovery index (CRI). Results showed that FPRs were between 0.05 and 0.075 under all conditions. For TPRs, magnitude had the largest effect among all design factors. With an ideal item bank, TPRs reached 0.8 when magnitude was 1.0. Scatter and shape had some effects when the directions of change were mixed (non-monotone) across testing occasions. In addition, Time 1 true θ value (θ_1) and its interactions had some effects under a practical item bank that had low test information at extreme θ values. CRIs were generally under 0.1 except at extreme θ_1 values, indicating good change recovery. The results showed that when the magnitude of change is large, AMC has sufficient power to detect and recover individual change, regardless of the scatter and the shape of that change. When the magnitude of change is small, significance testing results should be interpreted cautiously due to the lack of power.

Keywords: adaptive measurement of change, trait change patterns, computerized adaptive testing, longitudinal measurement

Change of an individual's trait level over time is known as intra-individual change. Measuring intra-individual change is important in psychological and educational assessments. For example, in hospital and clinical settings, clinicians are interested in whether a therapeutic intervention results in an improvement in psychological symptoms. Similarly, in school and training settings, instructors want to evaluate whether an instructional strategy results in an improvement in measured outcomes. These questions can only be answered by repeatedly measuring the same individual before and after the intervention and comparing the scores in a psychometrically sound manner.

Straightforward as it appears to be, measurement of individual change and testing its significance are challenging. Traditional approaches based on classical test theory (CTT) have been considered psychometrically flawed for more than 50 years (Cronbach & Furby, 1970). For instance, simply taking the difference between the pre- and post-intervention scores results in low reliability (Embretson, 1995; Hummel-Rossi & Weinberg, 1975; Lord, 1963; Willett, 1994, 1997), negative correlation with the initial status (Cronbach & Furby, 1970; Embretson, 1995; Willett, 1994, 1997), regression to the mean (Cronbach & Furby, 1970; Hummel-Rossi & Weinberg, 1975), and dependence on potentially different scales (Embretson, 1995; Hummel-Rossi & Weinberg, 1975). The reliable change index (RCI), used extensively in clinical applications (e.g., Jacobson et al., 1984; Jacobson & Truax, 1991; Marx et al., 2022), is defined as the ratio between a person's observed change score and the standard error of measurement (SEM) of their difference score. But the RCI has major drawbacks beyond those discussed above, including the use of a uniform SEM across all individuals under the CTT framework (e.g., Brouwer et al., 2013; Gulliksen, 1950; Lord & Novick, 1968) and arbitrariness in selecting the appropriate reliability index to compute the SEM (Wang et al., 2021). The RCI is also limited to using simple difference scores based on two testing occasions.

Most modern statistical procedures that model change, such as latent growth curve models (Bryk & Raudenbush, 1987; Grimm et al., 2017; Meredith & Tisak, 1990) and structural equation models (Bollen & Curran, 2006; Grimm et al., 2017), analyze change at the group level. When individual change is addressed, it is always considered in relation to group change. Group-level analysis is appropriate for evaluating whether an intervention is effective for a group, but inappropriate for evaluating whether change has occurred at the individual level (Wang & Weiss, 2018).

Adaptive measurement of change (AMC) represents a modern, psychometrically rigorous approach to measuring and testing the psychometric significance of individual change. In this context, an individual change having psychometric significance means that a null hypothesis significance testing procedure, based entirely on the individual's trait estimates and related data, determines that the trait level of the individual has changed over two or more testing occasions. First proposed by Weiss and Kingsbury (1984) under the name of "adaptive self-referenced testing," AMC integrates item response theory (IRT) and computerized adaptive testing (CAT) into a coherent psychometric procedure. In its initial conceptualization, two sets of CATs are administered at pre- and post-intervention occasions to estimate an examinee's trait (θ) values along with their SEMs, which in IRT will vary with θ levels. Then the confidence intervals of θ at the two occasions are constructed from the SEMs and compared. If the intervals do not overlap, psychometrically significant individual change is said to have occurred.

In recent years, several studies have extended the pioneering work of Weiss and Kingsbury (1984). For example, Kim-Kang and Weiss (2008) compared the performance of AMC with three conventional testing methods for measuring individual change across two occasions. They concluded that AMC outperformed all three CTT methods in the examined conditions. Not only did AMC

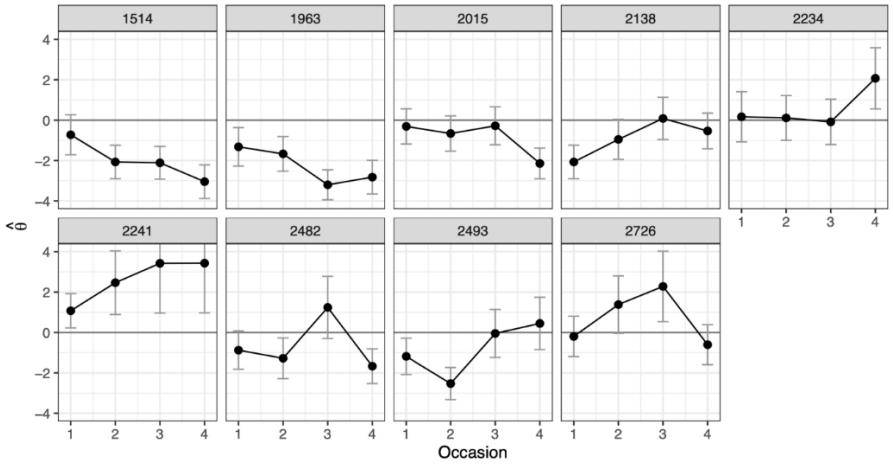
measure individual change equally well across the entire range of θ, it also dramatically reduced the number of items necessary to administer to detect significant change. Finkelman et al. (2010) enriched the AMC framework by investigating the performance of a new item selection procedure, namely Kullback–Leibler information (Cover & Thomas, 1991), and two new hypothesis testing methods—a Z-test and a likelihood-ratio chi-square test. They found improvements in the procedure by showing better adherence to Type I error rates and better power for detecting small changes. Lee (2015) further investigated three additional hypothesis testing methods and one new item selection method, concluding that the Z statistic displayed a better balance between Type I error rates and power compared to the other statistics.

Phadke (2017) examined the performance of AMC across more than two occasions. She developed and compared several generalized methods for hypothesis testing and concluded that the likelihood-ratio test achieved the optimum balance between Type I error and power. Wang and Weiss (2018) then extended the AMC framework to study multidimensional latent traits using multidimensional IRT under two testing occasions. Wang et al. (2021) further developed and investigated the performance of AMC methods for multidimensional latent traits measured across multiple occasions, which is the most comprehensive AMC framework. Cooperman et al. (2021) showed that AMC is generally robust to item parameter estimation error.

A significant limitation of the previous studies is that the effects of trait change patterns on AMC performance were not systematically investigated. The change pattern of an individual refers to the numerical characteristics of the changes over multiple testing occasions, including their magnitudes and directions (increase or decrease). Because most of the previous studies focused on hypothesis testing methods and the magnitude of change as the primary design factors, they did not design the change patterns systematically. Their common practice was to specify a limited number of change patterns that represented different arbitrary combinations of magnitude, direction, and linear or nonlinear patterns of change. However, Cooperman et al. (2021) showed that among design factors including hypothesis testing method, item parameter estimation error, starting θ value, and change pattern, the effect sizes of change pattern on most AMC performance indicators were far greater than other design factors. For example, in their results, the classical effect sizes (η^2) for true positive rates (TPRs, or power) were 0.793 for change pattern, 0.06 for starting θ, and 0.049 for the change pattern × starting θ interaction. As a comparison, hypothesis testing method had an effect size of only 0.021.

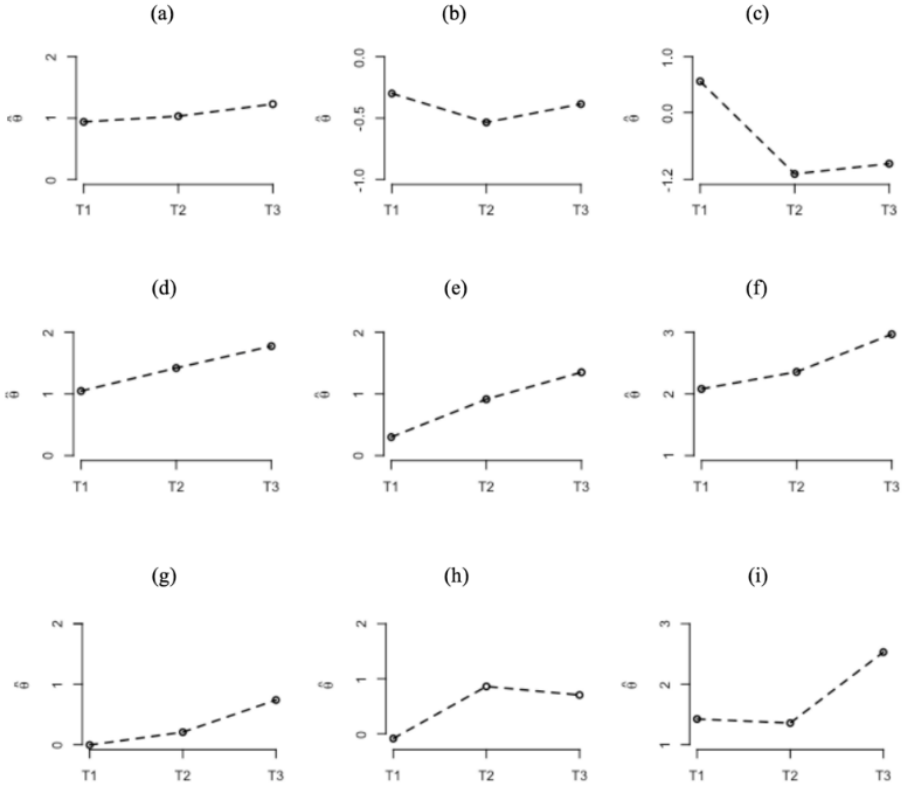
In reality, change patterns vary dramatically across individuals. As two illustrations, Figure 1 depicts the change patterns of the Applied Cognition scale scores from the patient-reported outcomes of nine hospitalized patients on four measurement occasions (Wang et al., 2022; Weiss et al., 2021); Figure 2 depicts the change patterns of the math ability of nine students on three testing occasions in a K–12 setting (Phadke, 2017). These patterns vary in terms of their total change magnitude, the dispersion of change across occasions, and the directions of change. These factors can have substantial impact on the performance of AMC. For example, Finkelman et al. (2010) showed that in the case of two testing occasions, TPRs could be as low as around 0.4 when the magnitude of change was 0.5 and as high as nearly 1.0 when the magnitude of change was 1.5. Cooperman et al. (2021) showed that in the case of four testing occasions, change patterns [+0.25, +0.25, +0.25], [+0.5, +0.25, 0], and [+0.75, 0, 0] resulted in TPRs of around 0.55, 0.60, and 0.65, respectively; in other words, patterns with the same total magnitude of change but different distributive patterns resulted in different TPRs. It is, therefore, important to investigate how each factor, as well as its interactions, impacts the ability of AMC to detect and recover change.

Figure 1. Applied Cognition θ Estimates and 2 SEM Bands for Nine Hospitalized Patients at Four Test Occasions



Note. The header of each panel is an arbitrary patient ID.

Figure 2. Mathematics Ability θ Estimates of Nine Students at Three Test Occasions



Purpose

The purpose of this study was to systematically specify the change patterns of latent traits and to investigate their effects on the performance of AMC. Specifically, change patterns were based on a framework proposed by Cronbach and Gleser (1953) that decomposes a profile pattern into three components: magnitude, scatter, and shape. The primary research question motivating this research was how each pattern component and their interactions affected the performance of AMC in terms of detecting change, if any, and estimating the magnitude and direction of the change. A monte-carlo simulation study was designed to examine how TPRs, false positive rates (FPRs; i.e., Type 1 error rates), and a change recovery index (CRI) were affected by three design factors: (1) trait change pattern; (2) starting θ value (θ_1); and (3) item bank characteristics, which are further described below.

Method

Design Factors

Change Patterns

Based on the approach developed by Cronbach and Gleser (1953), change patterns were characterized by three independent components: (1) magnitude, (2) scatter, and (3) shape. The shape component was further decomposed into two subcomponents, named direction and order. Change was simulated to have occurred across four testing occasions on a single variable.

Magnitude

Magnitude describes the total amount of change in $\boldsymbol{\theta}$ across all testing occasions. It can be defined as

$$Magnitude = \sum_{i=2}^{n} |\Delta\theta_i|, \tag{1}$$

where $\Delta\theta_i = \theta_i - \theta_{i-1}$ (i.e., the change in θ from occasion $i-1$ to i) and n is the number of testing occasions. Note that the summation is over the absolute value of $\Delta\theta_i$. As a numerical example, consider two patterns—Pattern A and Pattern B—as displayed in Figure 3.

Figure 3 displays the patterns in two formats: in θ units (Figure 3a) and in $\Delta\theta$ units (Figure 3b) (all subsequent figures in this section are presented in the same manner). Figure 3b shows that for Pattern A, $\Delta\theta_2 = \Delta\theta_3 = \Delta\theta_4 = 0.2$, so $\text{magnitude}_A = |\Delta\theta_1| + |\Delta\theta_2| + |\Delta\theta_3| = 0.6$. For Pattern B, $\Delta\theta_2 = \Delta\theta_3 = \Delta\theta_4 = 0.1$, so $\text{magnitude}_B = |\Delta\theta_1| + |\Delta\theta_2| + |\Delta\theta_3| = 0.3$. Therefore, Pattern A has larger magnitude than Pattern B.

Scatter

Scatter describes the dispersion of change in $\boldsymbol{\theta}$ ($\Delta\boldsymbol{\theta}$) across all testing occasions. It can be defined as

$$Scatter = \sqrt{\sum_{i=2}^{n}\left(|\Delta\theta_i| - \overline{|\Delta\theta|}\right)^2}, \quad (2)$$

where $\overline{|\Delta\theta|}$ is the average of $|\Delta\theta|$ across all testing occasions. For the two patterns displayed in Figure 3, both have zero scatter because their $\Delta\theta$s remain the same across occasions. Figure 4 displays two alternative patterns with different amounts of scatter.

Figure 3. Magnitude: Comparison of Pattern A and Pattern B

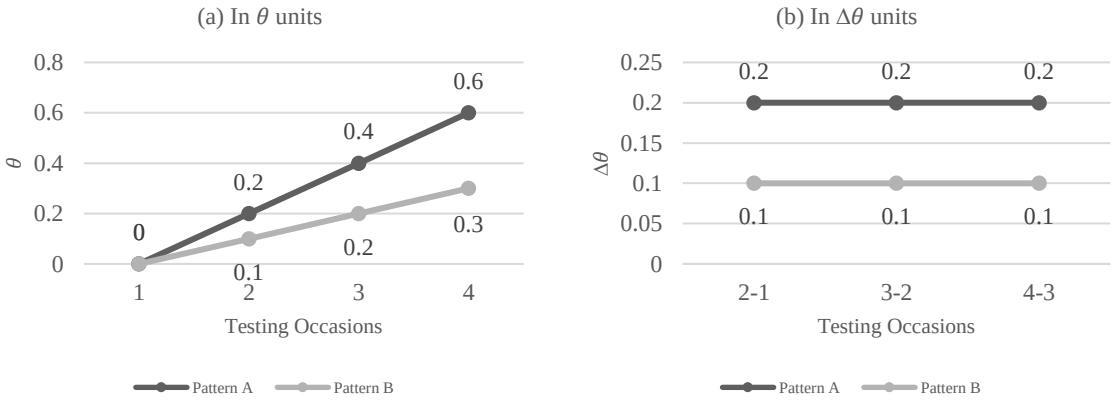


Figure 4. Scatter: Comparison of Pattern C and Pattern D

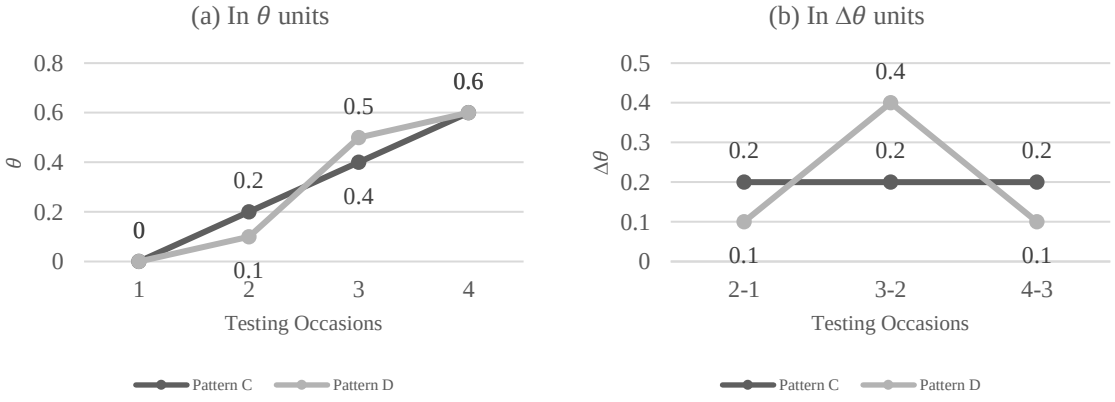


Figure 4b shows that for Pattern C, $\Delta\theta_2 = \Delta\theta_3 = \Delta\theta_4 = 0.2$, so $\overline{|\Delta\theta|}_c = 0.2$ and $scatter_C = 0$ by Equation 2. For Pattern D, $\Delta\theta_2 = 0.1$, $\Delta\theta_3 = 0.4$, $\Delta\theta_4 = 0.1$, so

$$\overline{|\Delta\theta|}_D = \frac{(|\Delta\theta_2| + |\Delta\theta_3| + |\Delta\theta_4|)}{3} = \frac{0.1 + 0.4 + 0.1}{3} = 0.2, \quad (3)$$

and,

$$scatter_D = \sqrt{\sum_{i=2}^{n}\left(|\Delta\theta_i| - \overline{|\Delta\theta|}\right)^2} = \sqrt{(0.1-0.2)^2 + (0.4-0.2)^2 + (0.1-0.2)^2} = 0.25. \quad (4)$$

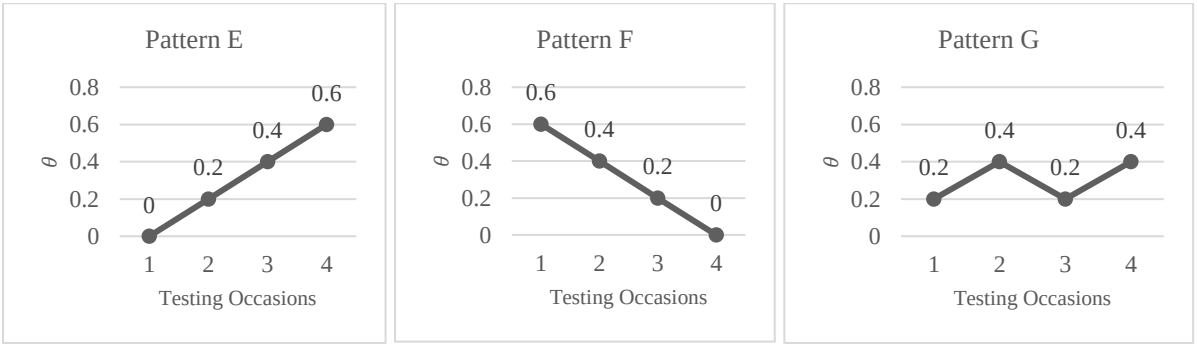
Therefore, Pattern D has larger scatter than Pattern C, even though they have the same magnitude of change (0.6).

Shape

Following the definition in Cronbach and Gleser (1953), shape contains the information about a pattern after accounting for its magnitude and scatter. In this study, shape information was further decomposed into direction and order. First, direction describes whether $\Delta\boldsymbol{\theta} > \mathbf{0}$ or $\Delta\boldsymbol{\theta} < \mathbf{0}$ (i.e., whether $\boldsymbol{\theta}$ is increasing or decreasing at a particular occasion). Consider the three patterns in Figure 5.

Figure 5. Direction: Comparison of Pattern E, Pattern F, and Pattern G

(a) In θ units



(b) In $\Delta\theta$ units

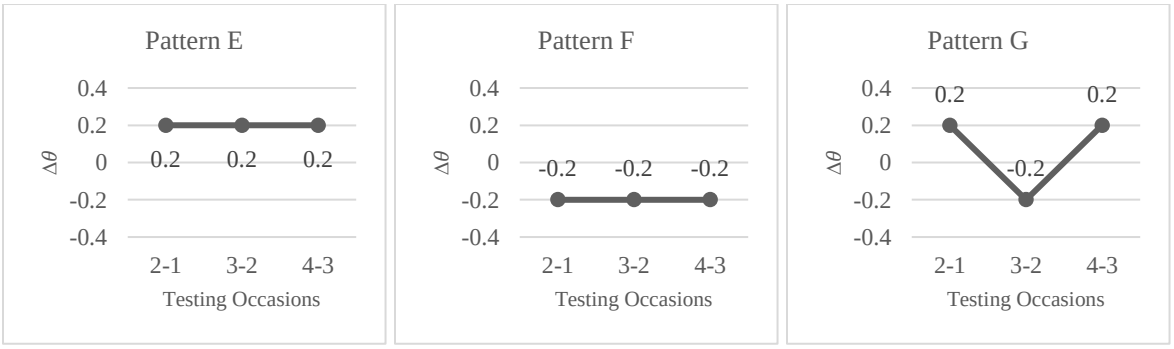


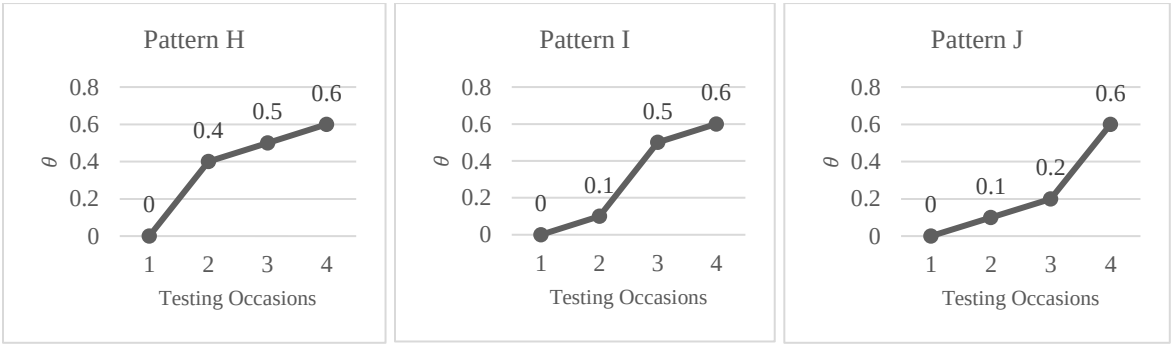
Figure 5b shows that for all three patterns, $|\Delta\theta_2| = |\Delta\theta_3| = |\Delta\theta_4| = \overline{|\Delta\theta|} = 0.2$. Therefore, they have the same amount of magnitude (0.6) and scatter (0). However, their directions of change are different across occasions. Using + and – to represent increase and decrease in θ, the three patterns can be represented as (+, +, +), (–,–,–), and (+,–, +), respectively.

Second, order describes the combinatorial order of $\Delta\theta$s. Consider the three patterns in Figure 6.

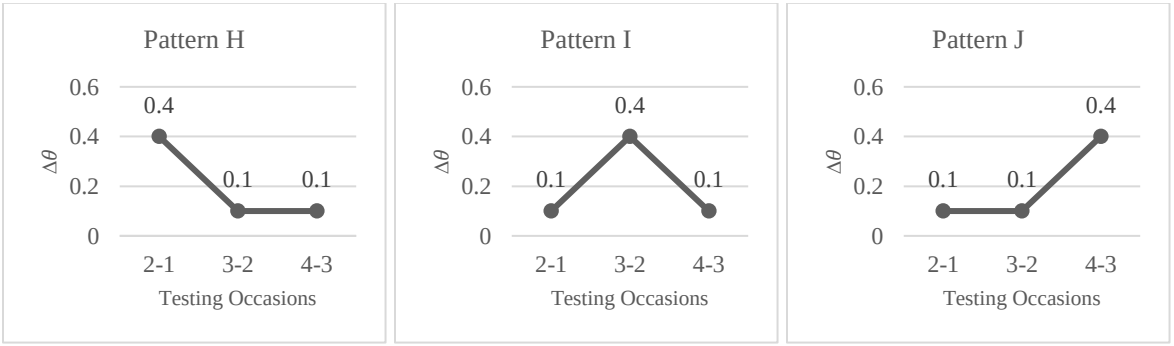
As Figure 6b shows, for all three patterns, the three $\Delta\theta$s at Occasions 2 to 4 are 0.1, 0.1, and 0.4. However, their combinatorial orders are [0.4, 0.1, 0.1], [0.1, 0.4, 0.1], and [0.1, 0.1, 0.4], respectively, where the brackets represent ordered combinations of elements. Otherwise, all three patterns have the same magnitude (0.6), scatter (0.25), and direction (+, +, +). Taken together, magnitude, scatter, and shape (including direction and order) describe the key information about a pattern in four independent components.

Figure 6. Order: Comparison of Pattern H, Pattern I, and Pattern J

(a) In θ units



(b) In $\Delta\theta$ units



Simulation Design

Four testing occasions were simulated for each simulee. Three levels of magnitude were considered, m = {0.5, 1.0, 1.5}. These levels were used in previous studies and considered as small, medium, and large changes, respectively (Finkelman et al., 2010; Lee, 2015; Phadke, 2017). In addition, they were based on the mean change scores in the classical percent-correct metric from educational data in the United States (Lee, 2015). The various patterns and types of change simulated were based on observed patterns of change on mathematics tests administered to K–12 students in the United States on three occasions (Phadke, 2017) and patient-reported outcomes data from hospitalized patients who were tested on from three to five testing occasions (Weiss et al., 2021), as illustrated in Figures 1 and 2.

Three scatter patterns were examined: [0.8m, 0.1m, 0.1m], [0.6m, 0.2m, 0.2m], and [0.4m,

0.3m, 0.3m]. These patterns corresponded to three scatter values: 0.43m, 0.29m, and 0.14m, respectively, which were reasonably spread out to allow its effects to be investigated. Note that all three patterns were in the form of [a, b, b], meaning that two of the three values were the same (denoted as b), whereas the third was different (denoted as a). Adding variations on order, three change patterns were possible: [a, b, b], [b, a, b], and [b, b, a]. Such a design reduced the number of simulations needed. For example, for the change pattern [a, b, b] there are only three possible orders: [a, b, b], [b, a, b], and [b, b, a], whereas for the change pattern [a, b, c] there are six possible orders. However, the reduction in the number of simulations did not impede investigating the impact of scatter on the performance of AMC.

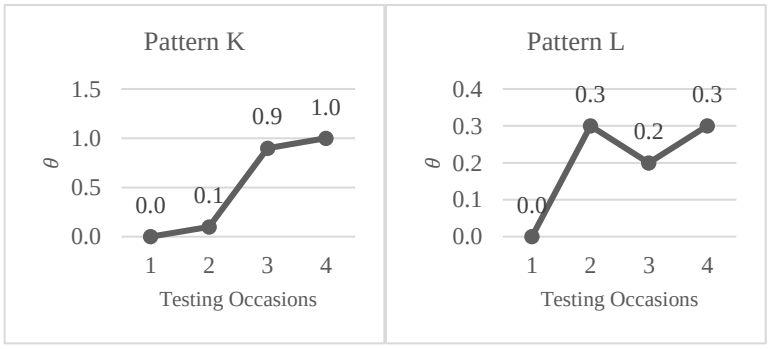
There were two cases for direction: unidirectional changes and mixed-directional changes. In the case of unidirectional changes, all $\Delta\theta$s were either positive or negative, meaning that the change pattern was either monotone increasing (+, +, +) or monotone decreasing (–, –, –). Adding variations in order, there were three possible change patterns for the monotone increasing case: [+a, +b, +b], [+b, +a, +b], and [+b, +b, +a], and three for the monotone decreasing case: [–a, –b, –b], [–b, –a, –b], and [–b, –b, –a].

In the case of mixed-directional changes, not all $\Delta\theta$s were positive or negative (i.e., the change pattern was mixed). In this case, for each of the three order conditions, there were six possible patterns. For example, for order condition (b, a, b), the six possible patterns were: [+b, +a, –b], [+b, –a, +b], [+b, –a, –b], [–b, +a, –b], [–b, –a, +b], and [–b, +a, +b].

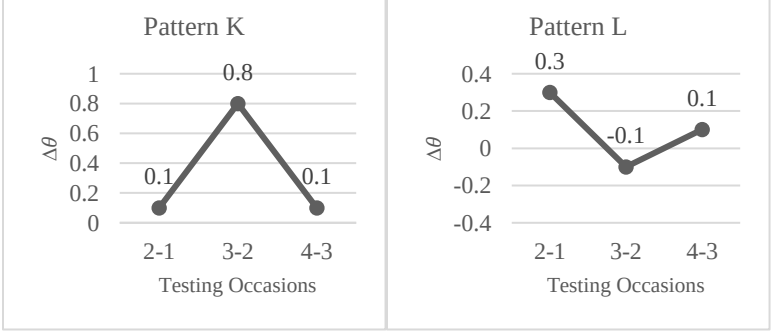
Two patterns are presented as examples in Figure 7. Pattern K is a unidirectional change case:

Figure 7. Examples: Patterns K and L

(a) In θ units



(b) In $\Delta\theta$ units



$\boldsymbol{\theta} = [0.0, 0.1, 0.9, 1.0]$, $\Delta\boldsymbol{\theta} = [0.1, 0.8, 0.1]$, magnitude m = 1.0, scatter = 0.43 (= 0.43m = 0.43 × 1.0 = 0.43), directions of change = [+, +, +], order = [b, a, b]. Pattern L is a mixed-directional change case, $\boldsymbol{\theta} = [0.0, 0.3, 0.2, 0.3]$, $\Delta\boldsymbol{\theta} = [0.3, -0.1, 0.1]$, magnitude m = 0.5, 0.29m = 0.29 × 0.5 = 0.145), directions of change = [+,–, +], order = [b, a, b], scatter = 0.145.

To summarize, Table 1 presents the number of change patterns across the design factors. There were 54 patterns in the case of unidirectional changes ($54 = 3 \times 3 \times 2 \times 3$) and 162 patterns in the case of mixed-directional changes ($162 = 3 \times 3 \times 6 \times 3$). In addition, a no-change pattern (i.e., $\Delta\theta_2 = \Delta\theta_3 = \Delta\theta_4 = 0$) was also examined to evaluate the FPRs, which should be around the prespecified α level (0.05). Altogether, the total number of change patterns examined was 217.

Table 1. Number of θ Patterns in the Simulation Plan

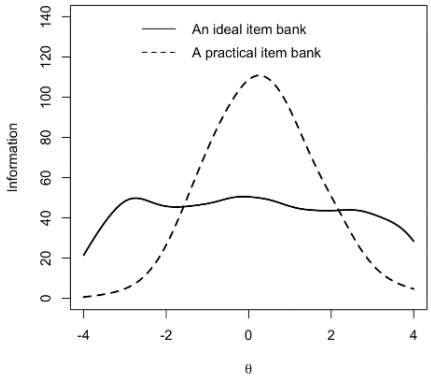
			Shape		
Type of change	Magnitude	Scatter	Direction	Order	Total
Unidirectional changes	3	3	2	3	54
Mixed-directional changes	3	3	6	3	162
No change					1
Grand total					217

θ_1 Values and Item Banks

Cooperman et al. (2021) showed that Time 1 true θ value (θ_1) and its interaction with change pattern had moderate effects on TPRs. They also found that item bank information played a substantial role in AMC performance, resulting in lower TPRs and more biased change recovery at extreme θ_1 values where bank information was low (also see Finkelman et al., 2010). Therefore, both θ_1 and item bank design were included as design factors. Following previous studies (Finkelman et al., 2010; Lee, 2015; Phadke, 2017), five discrete θ_1 values were chosen: $\theta_1 = \{-2, -1, 0, 1, 2\}$.

Two item banks were generated with 300 dichotomous items from the unidimensional three-parameter logistic (3PL) model (Birnbaum, 1986) with $D = 1.7$. The two banks shared the same set of discrimination parameters and pseudo-guessing parameters, but differed in their difficulty parameters. Specifically, the discrimination parameters were generated from a truncated normal distribution, N(1.25, 0.25), with bounds set at 0.50 and 2.0. A truncated normal distribution provides a good approximation of the discrimination parameters in a typical CAT item bank, where items with extreme discrimination parameters are often removed (Crichton, 1981). The pseudo-guessing parameters were fixed at 0.2 for all items (Lord & Novick, 1968). The difficulty parameters for the first item bank were generated from a uniform distribution, U[−4, 4], whereas those for the second item bank were generated from a normal distribution, N(0, 1.2) (Phadke, 2017). The first item bank therefore represents an item bank with equal measurement precision across the entire θ continuum, which is an ideal item bank for CAT (Kim-Kang & Weiss, 2008) because it has the potential to provide equiprecise measurement across θ. The second item bank represents a practical item bank with higher information around the center of the θ continuum and lower information at the two extremes, which resembles the shape of most real CAT item banks. These two banks were similar to those used in Finkelman et al. (2010). The bank information functions of the two banks are shown in Figure 8.

Figure 8. Item Bank Information Functions



CAT Administration

The AMC procedure aims to detect changes in estimated θ ($\hat{\theta}$) across the four testing occasions. Therefore, four CAT administrations were implemented, corresponding to the four occasions. The CAT starting value for Occasion 1 was 0.0 for all simulees. At Occasions 2, 3, and 4 (T_2, T_3, and T_4), the CAT starting values were the final ($\hat{\theta}$) from the previous occasion. For example, the CAT starting value at T_3 was $\hat{\theta}_2$, which was the final $\hat{\theta}$ from T_2. Each simulation condition was replicated 1,000 times to represent 1,000 simulees for each combination of pattern, θ_1 value, and CAT item bank. In addition, true (error-free) item parameters were used in the simulations because Cooperman et al. (2021) showed that item parameter estimation error had a small to negligible effect on AMC performance in similar testing scenarios.

All four testing occasions used fixed-length tests. To identify the optimal test length, a preliminary simulation study was conducted. The study results showed that given the two item banks described above, $\hat{\theta}$ stabilized after 25 items under all true θ conditions. Therefore, all four CAT administrations were administered as 25-item fixed-length tests. Throughout the AMC process, θ was estimated with maximum likelihood estimation (MLE). In case of the absence of a mixed-response vector, maximum a posteriori (MAP) was used with a standard normal prior distribution until a mixed-response pattern was obtained. Items were chosen to maximize the expected Fisher information conditional on the current $\hat{\theta}$ (Embretson & Reise, 2000).

Hypothesis Testing Method

Three hypothesis testing methods have been investigated in previous AMC research: Z-test, likelihood ratio test (LRT), and score ratio test. All three tests demonstrated desirable statistical properties in AMC (Finkelman et al., 2010; Lee, 2015; Phadke, 2017; Wang et al., 2021), except that the Z-test tended to perform worse than the other two at extreme θ_1 values (Cooperman et al., 2021). Therefore, the LRT was chosen for this study. Under the AMC context, the LRT statistic is the ratio of the likelihood of observing the response patterns under the null hypothesis ($H_0: \theta_1 = \theta_2 = \cdots = \theta_t$) over the likelihood of observing the response patterns under the alternative hypothesis (H_a: at least one of the equal signs does not hold), the latter being the product of the separate likelihoods evaluated at the corresponding θ estimate (Finkelman et al., 2010; Phadke, 2017). Formally, the LRT test statistic is defined as

$$\Lambda_O = \frac{L\left(\boldsymbol{u}_{1+2+\cdots+t}\middle|\hat{\theta}_{\text{Pool}_t}\right)}{L\left(\boldsymbol{u}_1\middle|\hat{\theta}_1\right) \times L\left(\boldsymbol{u}_2\middle|\hat{\theta}_2\right) \times \ldots \times L\left(\boldsymbol{u}_t\middle|\hat{\theta}_t\right)}, \tag{5}$$

where $\hat{\theta}_{\text{Pool}_t}$ is the MLE of θ under H_0, $\boldsymbol{u}_i$ is the response vector at testing occasion T_i $(i = 1, \ldots, t)$, $\boldsymbol{u}_{1+2+\cdots+t}$ is the combined response vector across t testing occasions, and $L(\cdot)$ is the likelihood value for the 3PL model evaluated at a $\hat{\theta}$ value. Under the null hypothesis, $-2\log_e \Lambda_O$ follows a chi-square distribution with $(t-1)$ degrees of freedom.

Dependent Variables

The performance of AMC was evaluated on three dependent variables: (1) TPR, (2) FPR, and (3) CRI. In the case of no change over time (i.e., $\Delta\theta_2 = \Delta\theta_3 = \Delta\theta_4 = 0$), FPRs measured the proportion of simulees that were identified by the AMC procedure as having significant change. TPRs measured the same proportion in cases where there was indeed change over time.

CRI was defined as

$$CRI = \sqrt{\left(\widehat{\Delta\theta_2} - \Delta\theta_2\right)^2 + \left(\widehat{\Delta\theta_3} - \Delta\theta_3\right)^2 + \left(\widehat{\Delta\theta_4} - \Delta\theta_4\right)^2}, \tag{6}$$

where $\widehat{\Delta\theta_2} = \widehat{\theta_2} - \widehat{\theta_1}$ (i.e., the estimated change in θ from T_1 to T_2) and $\Delta\theta_2 = \theta_2 - \theta_1$ (i.e., the true change in θ from T_1 to T_2). Other terms were defined similarly. CRI measures the performance of AMC in terms of how accurately it recovered true change.

A series of three-way analyses of variance (ANOVAs) examined the effects of the five design factors, including four factors in change patterns (magnitude, scatter, direction, and order) and θ_1 values, on each dependent variable. All factors were fully crossed. The classical effect sizes were computed as

$$\eta^2 = \frac{\text{SS}_{\text{Factor}}}{\text{SS}_{\text{Total}}}, \tag{7}$$

where SS denotes the sum-of-squares from each factor in the ANOVA. A separate ANOVA was run for each of the dependent variables.

Software

All simulations were conducted using R statistical software. The *catIrt* library (Nydick, 2014) was used to estimate θ and the *ggplot2* library (Wickham, 2016) was used to create plots. All other analyses were completed using functions written by the authors and adapted from Cooperman et al. (2021). The code is available upon request from the first author.

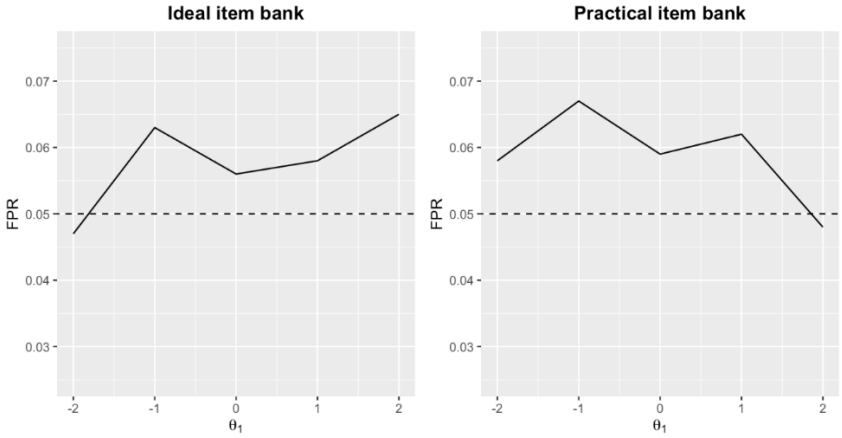
Results

FPR results are shown for two conditions: the practical item bank and the ideal item bank. TPR and CRI results are shown for four scenarios: unidirectional changes with a practical item bank (UDP), unidirectional changes with an ideal item bank (UDI), mixed-directional changes with a practical item bank (MDP), and mixed-directional changes with an ideal item bank (MDI).

FPRs

Under each combination of item bank and θ_1 conditions, the AMC procedure was replicated 1,000 times, and FPRs were calculated as the proportion of positive cases (i.e., cases when the null hypothesis was rejected by the LRT). Figure 9 shows that regardless of item bank and θ_1 value, FPRs hovered around 0.06, near the prespecified Type I error rate under the null hypothesis. No pattern was identified. The largest FPR was 0.067 at $\theta_1 = -1$ for the practical item bank. The simulations were repeated several times and similar results were observed. These results provided evidence that the AMC framework was performing as expected.

Figure 9. FPRs as a Function of Item Bank and θ_1



TPRs

Table 2 presents the three-way ANOVA results for TPRs, and Figure 10 shows how TPRs varied with magnitude, scatter, order, and θ_1. Similar to FPRs, the AMC procedure was replicated 1,000 times under each combination of magnitude, scatter, order, and θ_1 conditions, and TPRs were calculated as the proportion of positive cases. In the UDI condition (shown in the first column in Table 2 and in Figure 10a), magnitude had an effect size of 0.981. No other factors had an effect size of 0.01 (i.e., a small effect size; Cohen, 1988) or larger. Figure 10a shows that TPRs increased as magnitude increased (from left to right). Note that when magnitude was 1.0, TPRs were approximately 0.8 across the θ_1 continuum. TPRs were as low as between 0.25 and 0.3 for magnitude of 0.5 and nearly 1.0 for magnitude of 1.5. Direction of change had no effect, nor did scatter. Because order had negligible effect size, only order type (a,b,b) is shown in Figure 10.

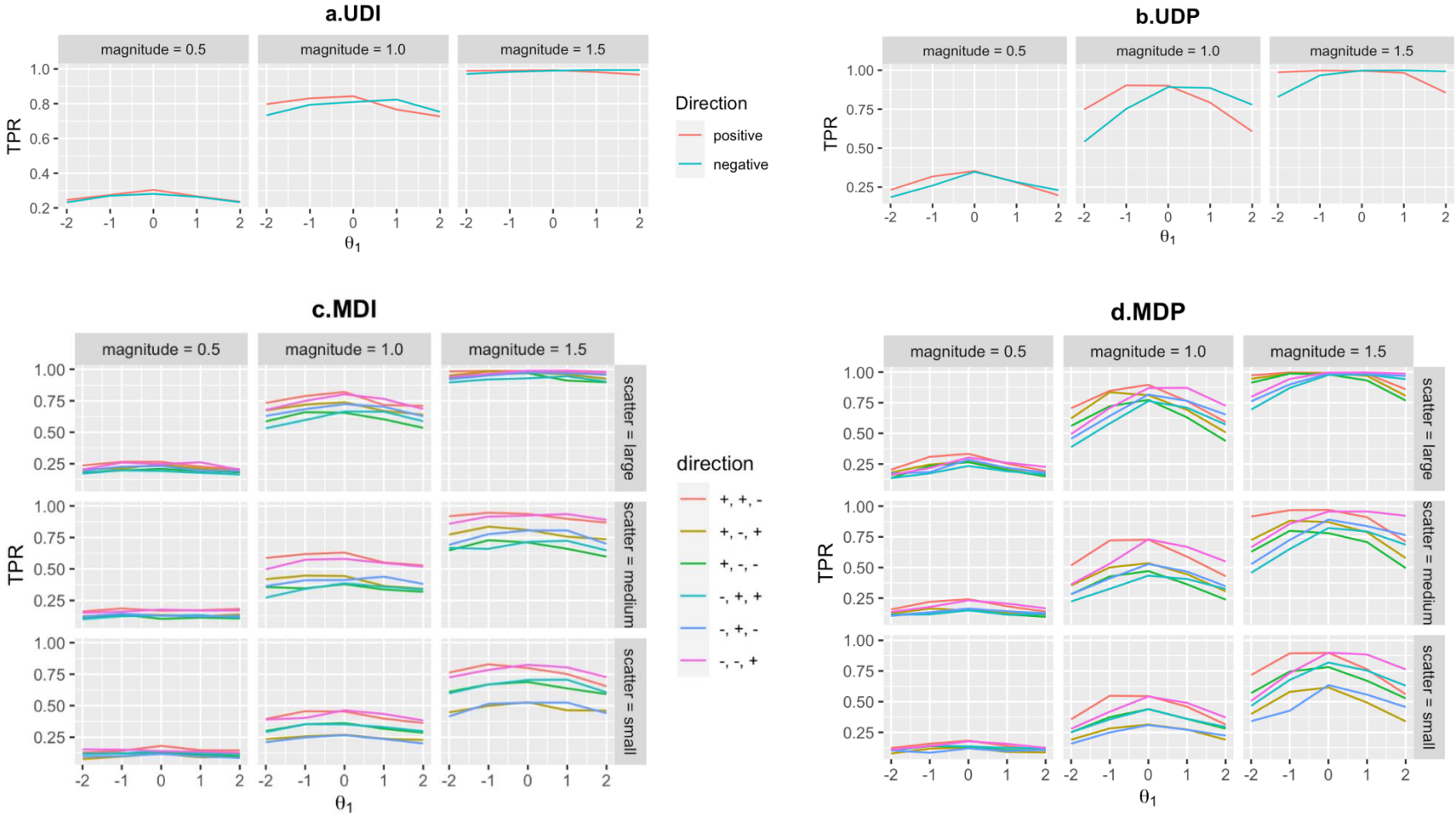
In the UDP condition (the second column in Table 2 and Figure 10b), the effect size of magnitude decreased to $\eta^2 = 0.909$. In this condition, θ_1 had an effect size of 0.039 and the θ_1 × direction interaction was 0.023. The main effect of θ_1 on TPRs can be observed from Figure 10b, where TPRs tended to be lower at the two extremes of θ_1 (ranging from 0.55 to 0.75) and higher in the middle (around 0.87). The θ_1 × direction interaction effect was evidenced by the positive and the negative direction line segments crossing each other, with the effect most pronounced when change magnitude was 1.0 and least evident with change magnitude of 0.5. When change magnitude was 1.0, TPRs of positive and negative direction at $\theta_1 = -2$ were 0.75 and 0.55, respectively; at $\theta_1 = 2$, they were 0.62 and 0.76, respectively. In the case of MDI (third column

Table 2. Classical Effect Sizes (η^2) From a Three-Way ANOVA on TPR for Each of the Four Change Scenarios

Factor	UDI	UDP	MDI	MDP
Main effects				
θ_1	0.003	0.039*	0.005	0.043*
magnitude	0.981***	0.909***	0.742***	0.700***
scatter	0.003	0.004	0.142***	0.129**
direction	0.000	0.000	0.020*	0.020*
order	0.004	0.004	0.006	0.005
Two-way interactions				
θ_1 × magnitude	0.001	0.006	0.001	0.007
θ_1 × scatter	0.000	0.000	0.000	0.000
θ_1 × direction	0.001	0.023*	0.000	0.005
θ_1 × order	0.000	0.000	0.000	0.000
magnitude × scatter	0.001	0.001	0.028*	0.023*
magnitude × direction	0.000	0.000	0.005	0.004
magnitude × order	0.001	0.001	0.001	0.001
scatter × direction	0.000	0.000	0.008	0.008
scatter × order	0.001	0.001	0.002	0.002
direction × order	0.000	0.000	0.021*	0.020*
Three-way interactions				
θ_1 × magnitude × scatter	0.000	0.000	0.001	0.003
θ_1 × magnitude × direction	0.001	0.007	0.000	0.002
θ_1 × magnitude × order	0.000	0.000	0.000	0.000
θ_1 × scatter × direction	0.000	0.000	0.000	0.001
θ_1 × scatter × order	0.000	0.000	0.000	0.000
θ_1 × direction × order	0.000	0.000	0.000	0.005
magnitude × scatter × direction	0.000	0.000	0.005	0.005
magnitude × scatter × order	0.000	0.000	0.001	0.001
magnitude × direction × order	0.000	0.000	0.004	0.003
scatter × direction × order	0.000	0.000	0.004	0.003
Residuals	0.001	0.001	0.004	0.009

* Indicates a small effect size (≥ 0.01). ** Indicates a medium effect size (≥ 0.06). *** Indicates a large effect size (≥ 0.14). These effect size thresholds are from Cohen (1988).

Figure 10. TPRs as a Function of Magnitude, Direction, θ_1, and Scatter for Unidimensional and Multidirectional Change (Order Type a, b, b)



in Table 2 and Figure 10c), the effect size of magnitude was $\eta^2 = 0.742$, still the largest among all factors. However, TPRs at any magnitude level varied noticeably with scatter. Compared to the UDI condition where scatter had negligible effect sizes, scatter in the MDI condition had an effect size of 0.142. This effect can be observed from Figure 10c. As scatter decreased (rows from top to bottom), TPRs decreased for all magnitudes of change. For instance, when magnitude was 1.0, TPRs were mostly in the range of 0.5 and 0.8 with large scatter, in the range of 0.3 and 0.6 with medium scatter, and in the range of 0.25 and 0.5 with small scatter.

In addition to magnitude and scatter, three other factors, namely direction, magnitude × scatter interaction, and direction × order interaction, each had an effect size of between 0.02 and 0.03. The effect of direction can be seen in Figure 10c, which shows that the direction patterns represented by the orange (+, +, –) and pink (–,–, +) line segments have higher TPRs than other direction patterns at all magnitude and scatter levels. The difference in TPRs due to different direction patterns could be as large as 0.3 under some large magnitude conditions. To see the magnitude × scatter interaction effect, note that although TPRs decreased as scatter decreased, the magnitude of TPR decrease increased as magnitude increased.

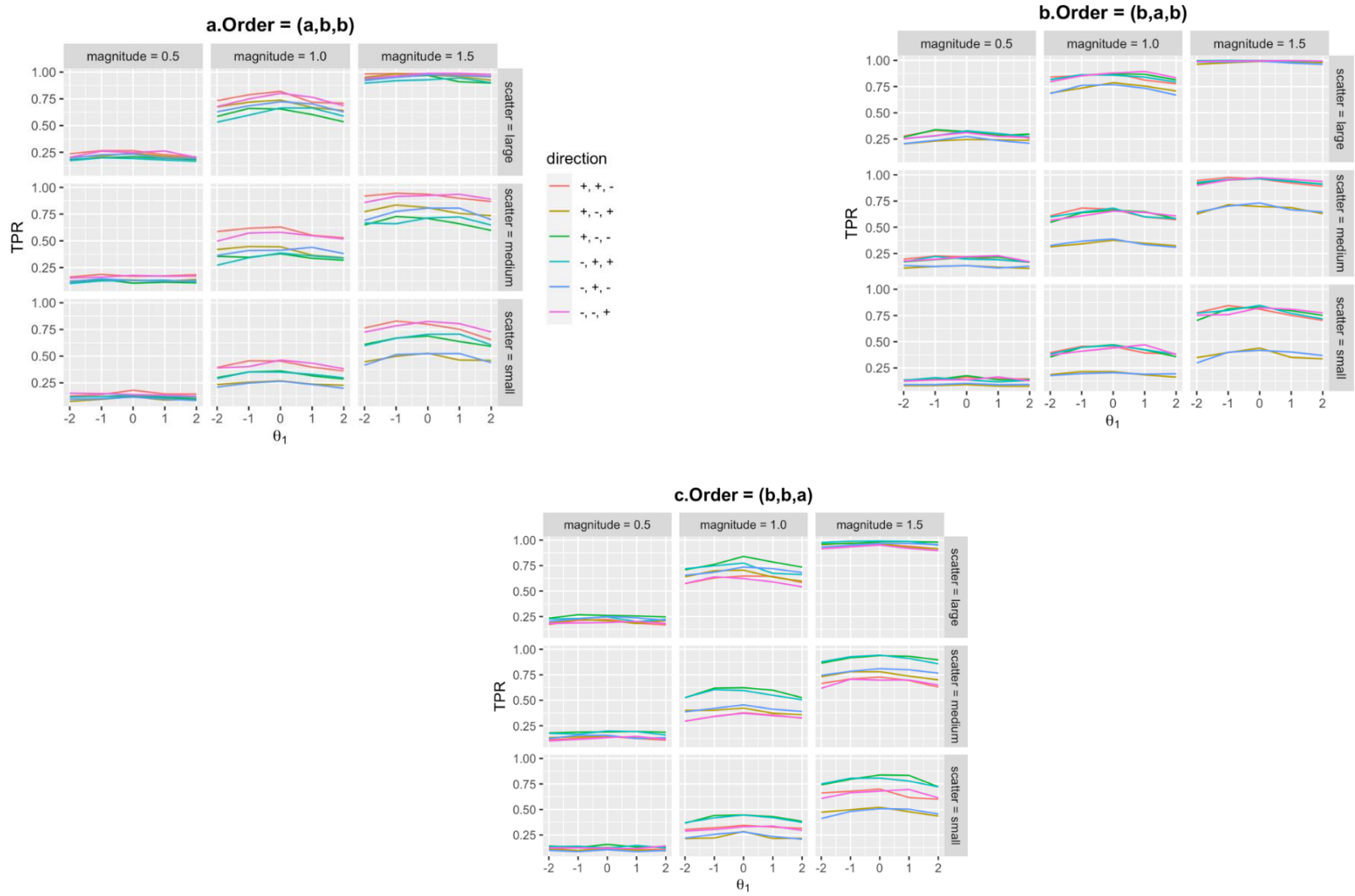
The direction × order interaction effect can be observed in Figure 11, which displays TPRs separately for three order conditions. Results in Figure 11a (order pattern a, b, b) are similar to those in Figure 10c, in which orange (+, +, – and pink (–, –, +) line segments have the highest TPRs. As contrast, in Figure 11b (order pattern b, a, b), the blue (–, +, –) and gold (+,–, +) patterns have consistently lower TPRs than other patterns, whereas in Figure 11c (order pattern b, b, a), the light green (–, +, +) and dark green patterns (+,–,–) have almost consistently higher TPRs than other patterns.

In the MDP condition (the fourth column in Table 2 and Figure 10d), the effect sizes of factors were quite similar to those in the ideal item bank case, with the notable exception of θ_1. With the ideal item bank, θ_1 had negligible effect size; with the practical item bank, as expected, θ_1 had an effect size of 0.043. Such an effect can be observed from Figure 10d, where TPRs tended to be higher in the middle range of θ_1 and lower at the two extremes. The difference could be as large as 0.4 under some large magnitude conditions.

To summarize the TPR results, two salient patterns were observed: First, the effect sizes of θ_1 were smaller under the ideal item bank (UDI and MDI) than under the practical item bank (UDP and MDP). This was due to the fact that the test information of the practical item bank was higher in the middle of the θ continuum and lower at the two extremes, resulting in a reduction in TPRs as θ_1 deviated from 0. As a comparison, the test information of the ideal item bank was roughly equal throughout the entire θ continuum, resulting in TPRs that were basically equal across the θ_1 range.

Second, the effect sizes of the magnitude main effect were smaller under mixed-directional conditions (MDP and MDI) than under unidirectional conditions (UDP and UDI), and those of the scatter main effect were larger under mixed-directional conditions than under unidirectional conditions. The reason is hypothesized to be because with mixed-directional changes, the combination of scatter, direction, and order control the allocation of total magnitude across different occasions, resulting in positive changes and negative changes canceling each other. As a result, magnitude showed smaller effect sizes. To substantiate this hypothesis, magnitude was replaced by net magnitude and the data were reanalyzed. Net magnitude was defined as the absolute value of the difference in θ between the last occasion and the first occasion, that is,

Figure 11. TPRs Under MDI Conditions, By Order



$$Net\ Magnitude = |\theta_4 - \theta_1|. \quad (8)$$

The difference between magnitude (as originally defined) and net magnitude, with Patterns E, F, and G, are illustrated in Figure 5. As discussed previously, all three patterns have the same magnitude, which is 0.6. However, for Patterns E and F, their net magnitude is also 0.6 (Pattern E: $|0.6 - 0| = 0.6$; Pattern F: $|0 - 0.6| = 0.6$), whereas that of Pattern G is only 0.2 ($|0.2 - 0| = 0.2$). The net magnitude of Pattern G is smaller because its changes are mixed-directional (+0.2, –0.2, +0.2), resulting in the positive changes and the negative change cancelling each other. In other words, net magnitude partially absorbs the allocative effect described above. If the allocative effect was important, the effect sizes of magnitude should increase and those of non-magnitude factors should decrease.

The data were analyzed by ANOVA using net magnitude. The results are displayed in Table 3. Noticeably, the effect size of scatter sharply decreased to 0.01 from 0.142 and 0.129 for the ideal bank and the practical bank, respectively. On the other hand, the magnitude main effect exhibited some noticeable increases. These results provided some support for the hypothesis concerning the allocative effect under mixed-directional change conditions.

CRIs

Table 4 presents the three-way ANOVA results for CRIs, and Figure 12 shows how CRIs varied with magnitude, scatter, and θ_1. Note that for CRI, a small value reflects better recovery of change than a large value. With the ideal item bank (the first and the third columns in Table 4), the factor with the largest effect size was the residuals, with $\eta^2 = 0.450$ in the case of unidirectional changes and 0.533 in the case of mixed-directional changes. Other than the residuals, factors with a medium effect size or above were all three-way interactions. Observing Figure 12a and 12c, no clear patterns are discernible.

With the practical item bank (the second and the fourth columns in Table 4), magnitude, magnitude × θ_1, and magnitude × θ_1 × direction all had effect sizes that were close to or above medium size (0.06). In addition, θ_1 and θ_1× direction had effect sizes between medium and large (0.14). Figure 12b and 12d show that at extreme θ_1 values, CRIs were higher under some high-magnitude conditions at the extremes of θ_1.

Regarding the magnitude of CRI, with the ideal item bank, CRIs under all conditions were less than 0.1. Because there appears to be no previous literature that defined CRI in the same way as it was used here, there was no standard against which to compare the results. However, based on how CRI was defined, 0.1 appears to be a small magnitude given the range of θ considered, meaning that change recovery was satisfactory. With the practical item bank, CRIs under most conditions were also less than 0.1, indicating good recovery. The only exceptions were when θ_1 took on extreme values. This was expected because the test information at extreme θ values under the practical item bank was relatively low, resulting in less accurate estimated θ and change in θ.

Table 3. η^2 From a Three-Way ANOVA on TPRs for Mixed-Directional Change Conditions Under MDI and MDP Conditions, Based on Net Magnitude and Total Mmagnitude

	MDI		MDP	
Factor	Net magnitude	Total magnitude	Net magnitude	Total magnitude
Main effects				
θ_1	0.005	0.005	0.043*****	0.043*****
magnitude	0.815***	0.742***	0.755***	0.700***
scatter	0.010*	0.142***	0.010*	0.129**
direction	0.020*	0.020*	0.019*	0.020*
order	0.006	0.006	0.006	0.005
Two-way interactions				
θ_1 × magnitude	0.001	0.001	0.006	0.007
θ_1 × scatter	0.000	0.000	0.000	0.000
θ_1 × direction	0.000	0.000	0.005	0.005
θ_1 × order	0.000	0.000	0.000	0.000
magnitude × scatter	0.041*	0.028*	0.039*	0.023*
magnitude × direction	0.004	0.005	0.004	0.004
magnitude × order	0.005	0.001	0.005	0.001
scatter × direction	0.010*	0.008	0.009	0.008
scatter × order	0.002	0.002	0.001	0.002
direction × order	0.059*	0.021*	0.061**	0.020*
Three-way interactions				
θ_1 × magnitude × scatter	0.000	0.001	0.002	0.003
θ_1 × magnitude × direction	0.001	0.000	0.005	0.002
θ_1 × magnitude × order	0.000	0.000	0.000	0.000
θ_1 × scatter × direction	0.000	0.000	0.003	0.001
θ_1 × scatter × order	0.000	0.000	0.000	0.000
θ_1 × direction × order	0.000	0.000	0.003	0.005
magnitude × scatter × direction	0.002	0.005	0.002	0.005
magnitude × scatter × order	0.000	0.001	0.000	0.001
magnitude × direction × order	0.018*	0.004	0.016*	0.003
scatter × direction × order	0.000	0.004	0.000	0.003
Residuals	0.001	0.004	0.003	0.009

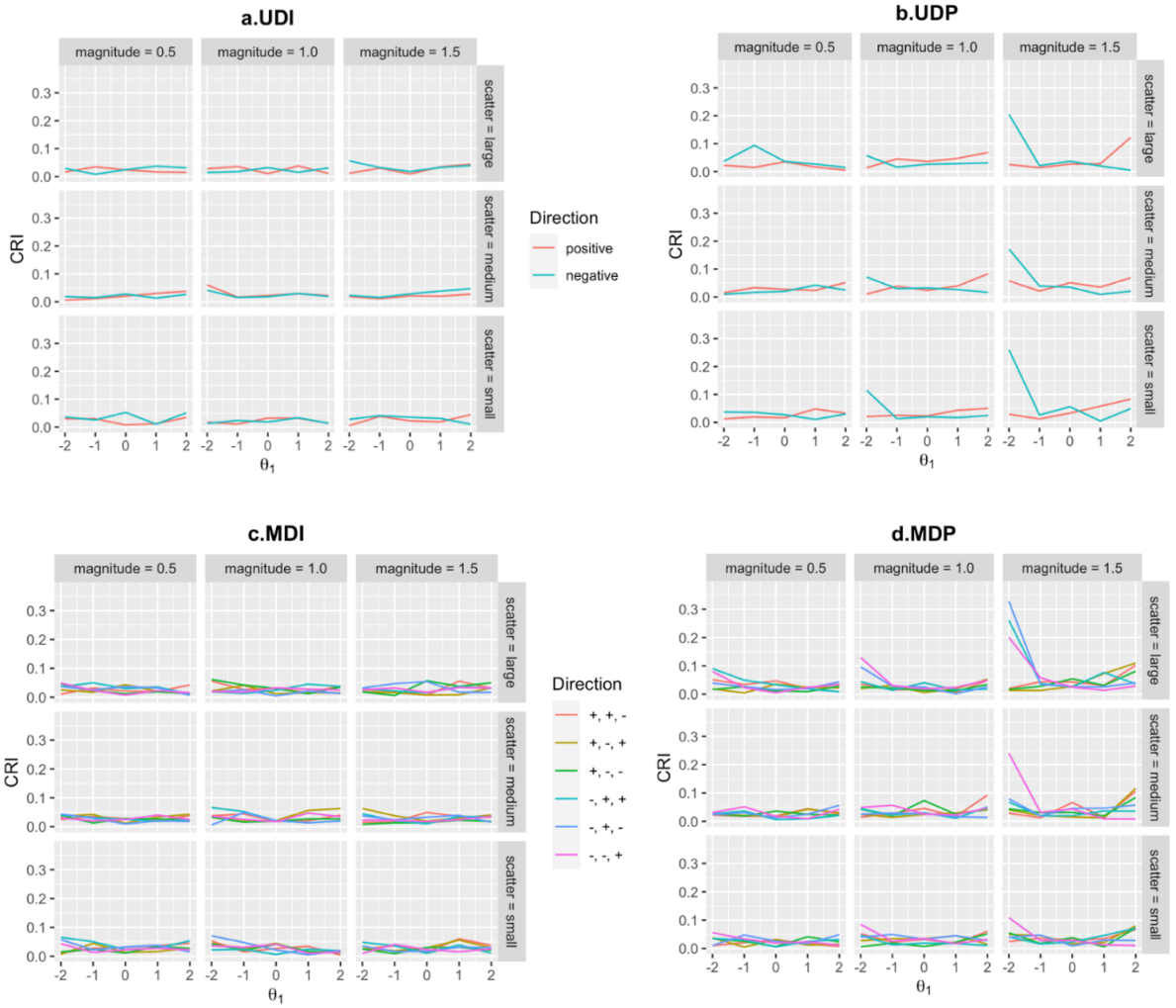
* Indicates a small effect size (≥ 0.01). ** Indicates a medium effect size (≥ 0.06). *** Indicates a large effect size (≥ 0.14).

Table 4. η^2 From a Three-Way ANOVA on CRI for the Four Item Bank Conditions

Factor	UDI	UDP	MDI	MDP
Main effects				
θ_1	0.038	0.172***	0.038	0.129**
magnitude	0.005	0.084**	0.004	0.051
scatter	0.002	0.001	0.003	0.013
direction	0.001	0.018	0.001	0.009
order	0.007	0.002	0.000	0.004
Two-way interactions				
θ_1 × magnitude	0.054	0.162***	0.008	0.055
θ_1 × scatter	0.030	0.004	0.017	0.027
θ_1 × direction	0.009	0.225***	0.031	0.095**
θ_1 × order	0.008	0.007	0.020	0.009
magnitude × scatter	0.004	0.002	0.002	0.008
magnitude × direction	0.005	0.017	0.017	0.004
magnitude × order	0.018	0.001	0.009	0.006
scatter × direction	0.004	0.000	0.010	0.012
scatter × order	0.004	0.003	0.008	0.001
direction × order	0.012	0.001	0.020	0.010
Three-way interactions				
θ_1 × magnitude × scatter	0.069**	0.007	0.020	0.021
θ_1 × magnitude × direction	0.057	0.188***	0.054	0.069**
θ_1 × magnitude × order	0.083**	0.007	0.017	0.005
θ_1 × scatter × direction	0.023	0.006	0.054	0.029
θ_1 × scatter × order	0.047	0.009	0.024	0.010
θ_1 × direction × order	0.024	0.002	0.040	0.082**
magnitude × scatter × direction	0.013	0.002	0.017	0.011
magnitude × scatter × order	0.014	0.001	0.007	0.004
magnitude × direction × order	0.008	0.003	0.015	0.010
scatter × direction × order	0.010	0.004	0.028	0.012
Residuals	0.450***	0.071**	0.533***	0.314***

* Indicates a small effect size (≥ 0.01).** Indicates a medium effect size (≥ 0.06). *** Indicates a large effect size (≥ 0.14).

Figure 12. CRI as a Function of Magnitude, Scatter, Direction, and θ_1



Discussion and Conclusions

This study investigated how the magnitude, the scatter, and the shape of a change pattern influenced AMC's performance in terms of detecting and recovering the pattern. FPRs under all conditions were between 0.05 and 0.075. Given that the pre-specified Type I error rate under the null hypothesis was 0.05, the results provided evidence that the AMC framework was performing as expected. There was a slight tendency for the practical bank to have slightly better FPRs than the ideal bank, which appears to be contrary to expectations—typically ideal item banks have better performance characteristics than actual/practical item banks (e.g., Finkelman et al., 2010; Lee, 2015; Phadke, 2017). However, in this case, the range of θ examined was between 2 and –2, and as Figure 8 shows, the practical item bank had considerably more information than the ideal item bank over the majority of that θ range. The performance of the AMC likelihood ratio test has been shown to be highly dependent on item bank information (Finkelman et al., 2010; Lee, 2015; Phadke, 2017).

The magnitude of change had the largest effect on AMC's ability to detect change among all design factors. Under mixed-directional change conditions, it had effect sizes greater than 0.7; under unidirectional change conditions, the effect sizes further increased to over 0.9. In order to reach TPRs of 0.8, magnitude needed to be 1.0 or higher. When magnitude was 0.5, TPRs were mostly between 0.1 and 0.4. When magnitude was 1.5, TPRs were close to 1.0, except for extreme θ_1 conditions. These results were broadly consistent with Finkelman et al. (2010), even though their study had only two testing occasions rather than four. The results were also consistent with the findings in Cooperman et al. (2021), which showed that with four testing occasions, a change magnitude of 0.75 resulted in TPRs of between 0.5 and 0.65. These results suggested that AMC might be underpowered when the magnitude of change is small, at least with the item banks studied thus far.

Scatter had almost no effect under unidirectional change conditions, but it had medium to large effects under mixed directional change conditions. As mentioned above, this might be due to the fact that scatter and shape controlled how total magnitude was allocated across testing occasions. Under unidirectional change conditions, TPRs were not affected because the total amount of change remained unchanged regardless of how it was distributed across occasions. But under mixed-directional change conditions, different combinations of scatter and shape resulted in different net total change from the first occasion to the last because positive changes and negative changes canceled each other. Indeed, when net magnitude replaced total magnitude in ANOVA, the effect size of scatter decreased drastically to 0.01.

Shape had two subcomponents: direction and order. Under unidirectional change conditions, neither had any salient effect. Under mixed-directional change conditions, direction and direction × order had small effects. Similar to scatter, these effects might be attributed to their allocative effects on magnitude. But unlike scatter, their effect sizes did not decrease when net magnitude replaced total magnitude in ANOVA. The effects of shape warrant further investigation in future studies.

θ_1 had small to medium effects with the practical item bank, but not with the ideal item bank. As discussed in the Results section, this could be explained by the different shapes of the item bank information functions. In addition, it can be observed from Figure 10b that for the practical item bank, at the higher end of θ_1 positive change resulted in lower TPRs than negative change,

whereas the opposite was true at the lower end of θ_1. In other words, there was a direction × θ_1 interaction. This is probably because when θ_1 was large (or small), positive (or negative) changes resulted in even more extreme θ values at subsequent occasions where test information was low, resulting in low TPRs.

Regarding CRIs, no clear patterns were identified for conditions with the ideal item bank. For conditions with the practical item bank, magnitude, direction, and θ_1 demonstrated main effects and interaction effects. As discussed previously, such effects were likely due to low test information at extreme θ_1 values. This can be observed from Figure 12b and 12d, where CRIs were higher at extreme θ_1 values (reflecting poorer recovery of change) for some high magnitude conditions. For example, as Figure 12b shows, CRIs were significantly higher when $\theta_1 = -2$, magnitude was 1.5, and the direction was negative, regardless of the level of scatter. In this case, θ values at subsequent testing occasions went further into the negative extreme where test information was low, resulting in poor estimation. Similarly, as the upper-right panel of Figure 12d shows, CRIs were significantly higher when $\theta_1 = -2$, magnitude was 1.5, scatter was large, and the direction of change from the first occasion to the second was negative (represented by the deep blue, light blue, and pink lines); these three change patterns were [−1.2, +0.15, −0.15], [−1.2, +0.15, +0.15], and [−1.2, −0.15, +0.15], respectively. In all three cases, $\theta_2 = -2 - 1.2 = -3.2$, where test information was low. Interestingly, higher CRIs, indicating poorer recovery of change, were mostly observed for negative values of θ_1 but not for positive values. This can be explained by the fact that the test information of the practical item bank was higher at $\theta_1 = 2$ than at $\theta_1 = -2$, as shown in Figure 8, due to the use of the 3PL model.

Comparing the results of the ideal item bank and the practical item bank, it appears that the crux of estimation performance as reflected in the CRI was bank information. The ideal item bank had approximately equal information across the $\theta_1 = [-4, 4]$ continuum, covering all possible θ values in this study. Therefore, CRIs under all conditions with the ideal item bank were low, and no design factor had a salient effect on CRI. On the other hand, the practical item bank had low information at the two extremes (especially the negative end). As a result, when θ on any occasion entered the low-information regions, estimation became imprecise, and CRI was high.

Limitations

A major limitation of this study is that the selection of magnitude values and scatter patterns was, to some extent, arbitrary (direction and order patterns were exhaustive, though). This was unavoidable because in order to perform ANOVA, each factor must have two or more levels. Given that magnitude and scatter were continuous variables, the only option was to specify several discrete values based on prior research and the authors' discretion. If different values and patterns had been used, the results might have been different. To investigate to what extent this issue would change the conclusions, another simulation study was performed with a different set of magnitude values (0.25, 0.50, and 0.75) and some other scatter patterns (for example, equal amount of change across occasions, i.e., zero scatter). Results showed that the patterns of ANOVA results were not fundamentally altered, despite slight changes in the numerical values. Therefore, it is expected that the main conclusions from the ANOVA results hold, despite the arbitrariness in assigning magnitude values and scatter patterns. Nevertheless, it will be helpful to replicate the study with different sets of magnitude values and scatter patterns. Another source of arbitrariness was the number of testing occasions, which was four in this study. Other numbers of testing occasions should be studied in the future to determine how AMC functions across larger numbers of testing occasions.

Another limitation was that only two item banks were used in this study, and they shared a common set of discrimination parameters. A wider range of item banks should be examined to gauge the generalizability of the current findings.

Conclusions

This study investigated the effects of change patterns on the performance of AMC. When there was no change, AMC performed satisfactorily, as FPRs were between 0.05 and 0.07 under all conditions. When there was change, the magnitude of change had the largest effect on the ability of AMC to detect change. In order to reach a TPR level of 0.8, change magnitude generally needed to be at least 1.0. When magnitude was small, AMC appeared to be underpowered. Scatter and shape each could change TPR by 0.3 under mixed-directional change conditions, but they had little effect under unidirectional change conditions. θ_1 and its interactions had some effects with the practical item bank due to low test information at extreme θ values, but they had no effect with the ideal item bank. CRI was generally lower than 0.1 except under some extreme θ_1 conditions, indicating good change recovery. In general, the results support the use of AMC for identifying psychometrically significant intra-individual change, especially when the magnitude of change is large. Still, it is important to note that insignificant testing results under AMC may be due to insufficient power, likely because of an item bank with low information, instead of lack of true change.

The results of this study have implications for the applications of AMC. Importantly, this study showed that AMC is robust to change patterns—its performance is primarily determined by the magnitude of change, not the scatter and the shape of the change pattern. This feature makes AMC an ideal framework for application because most practitioners are primarily concerned with the magnitude of change (e.g., has the student/patient improved? If so, how much?). In addition, when changes are mixed-directional across multiple occasions, practitioners are usually interested in comparing the trait level at the last occasion to the first occasion. This is exactly how net magnitude was defined in this study, and the results showed that net magnitude explained AMC performance even better than total magnitude. Moreover, the results using CRIs showed that change recovery was excellent under most conditions, meaning that AMC allows practitioners to obtain reliable estimates of change magnitude.

The power of AMC warrants further discussion. In the applications of AMC in educational contexts (Phadke, 2017) and clinical contexts (Weiss et al., 2021), one of their major limitations was that because the power of AMC was unknown, it was impossible to assess how many students/patients with true changes went undetected (i.e., cases of false negatives). This study showed that to achieve a power of 0.8, at least a change magnitude of 1.0 standard deviation of θ is necessary under an ideal item bank (i.e., one with a flat test information function); an even larger change magnitude is needed for a practical item bank (i.e., one with a peaked test information function). In addition, extreme starting θ values and mixed-directional changes lowered power. Therefore, it was likely that many students/patients with a smaller change magnitude went undetected in previous studies. Indeed, in its application to K–12 data, Phadke (2017) found that using the same LRT used in this study, only 25.3% of the students in the sample were detected as having change on their mathematics achievement over a full school year, which appears to be low. Based on the results of this study, it is possible that the true proportion of students with significant change was underestimated by AMC.

Consequently, it is recommended that AMC practitioners estimate the change in trait levels

(e.g., the difference between the last occasion and the first occasion) regardless of the hypothesis testing results. When hypothesis testing detects a change, it can be trusted with high confidence, because this study showed that the FPRs of AMC can be controlled near the nominal level, meaning that the probability of false positives is generally low and the estimated change in trait levels can be trusted. When AMC does not detect a change, the estimated change in trait levels can still provide useful information. Using practical judgment, test administrators can decide whether such a magnitude might indicate a change and, if it does, whether it has practical significance.

At a fundamental level, the power of AMC is strongly influenced by item bank characteristics, which was demonstrated by its performance at different trait levels under the practical item bank. To increase the power of AMC, it is suggested that test developers write items that are (1) informative (i.e., having high discriminations) and (2) well distributed in difficulty/location across the trait, which results in an item bank that provides more information at the extremes of the trait continuum. In addition, they can consider lengthening the CATs to improve the precision of trait estimation.

References

Birnbaum, A. (1986). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Addison-Wesley.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Wiley-Interscience.

Brouwer, D., Meijer, R. R., & Zevalkink, J. (2013). Measuring individual significant change on the Beck Depression Inventory-II through IRT-based statistics. *Psychotherapy Research, 23*(5), 489–501. CrossRef

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101*(1), 147–158. CrossRef

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Cooperman, A. W., Weiss, D. J., & Wang, C. (2021). Robustness of adaptive measurement of change to item parameter estimation error. *Educational and Psychological Measurement, 82*(4), 643–677. CrossRef

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. John Wiley.

Crichton, L. I. (1981). *Effect of error in item parameter estimates on adaptive testing* [Unpublished doctoral dissertation]. University of Minnesota, Twin Cities.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change." Or should we? *Psychological Bulletin, 74*(1), 68–80. CrossRef

Cronbach, L. J., & Gleser, G. C. (1953). Assessing similarity between profiles. *Psychological Bulletin, 50*(6), 456–473. CrossRef

Embretson, S. E. (1995). Measurement model for linking individual learning to processes and knowledge: Application to mathematical reasoning. *Journal of Educational Measurement, 32*(3), 277–294. CrossRef

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

Finkelman, M. D., Weiss, D. J., & Kim-Kang, G. (2010). Item selection and hypothesis testing for the adaptive measurement of change. *Applied Psychological Measurement, 34*(4), 238–254. CrossRef

Grimm, K., Ram, N., & Estabrook, R. (2017). *Growth modeling: Structural equation and multi-level modeling approaches*. Guilford.

Gulliksen, H. (1950). *Theory of mental tests*. Wiley.

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 (Ms. No. 916). *JSAS Catalog of Selected Documents in Psychology, 5*, 226.

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy, 15*(4), 336–352. CrossRef

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*(1), 12–19. CrossRef

Kim-Kang, G., & Weiss, D. J. (2008). Adaptive measurement of individual change. *Zeitschrift für Psychologie/Journal of Psychology, 216*(1), 49–58. CrossRef

Lee, J. E. (2015). *Hypothesis testing for adaptive measurement of individual change* [Unpublished doctoral dissertation]. University of Minnesota, Twin Cities.

Lord, F. M. (1963). Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 21–38). University of Wisconsin Press.

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

Marx, B. P., Lee, D. J., Norman, S. B., Bovin, M. J., Sloan, D. M., Weathers, F. W., Keane, T. M., & Schnurr, P. P. (2022). Reliable and clinically significant change in the clinician-administered PTSD Scale for DSM-5 and PTSD Checklist for DSM-5 among male veterans. *Psychological Assessment, 34*(2), 197–203. CrossRef

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika, 55*(1), 107–122. CrossRef

Nydick, S. (2014). *catIrt: An R package for simulating IRT-based computerized adaptive tests* (Version 0.50-0). WebLink

Phadke, C. (2017). *Measuring intra-individual change at two or more occasions with hypothesis testing methods* [Unpublished doctoral dissertaton]. University of Minnesota, Twin Cities.

Wang, C., & Weiss, D. J. (2018). Multivariate hypothesis testing methods for evaluating significant individual change. *Applied Psychological Measurement, 42*(3), 221–239. CrossRef

Wang, C., Weiss, D. J., & Suen, K. Y. (2021). Hypothesis testing methods for multivariate multi-occasion intra-individual change. *Multivariate Behavioral Research, 56*(3), 459–475. CrossRef

Wang, C., Weiss, D. J., Suen, K. Y., Basford, J., & Cheville, A. (2022). Multidimensional computerized adaptive testing: A potential path toward the efficient and precise assessment of applied cognition, daily activity, and mobility for hospitalized patients. *Archives of Physical Medicine and Rehabilitation, 103*(5), S3–S14. CrossRef

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*(4), 361–375. CrossRef

Weiss, D. J., Wang, C., Cheville, A., & Basford, J., & DeWeese, J., (2021). Adaptive measurement of change: A novel method to reduce respondent burden and detect significant individual-level change in patient-reported outcome measures. *Archives of Physical Medicine and Rehabilitation, 103*(5), S43–S52. CrossRef

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (2nd ed.). Springer. CrossRef

Willett, J. B. (1994). Measurement of change. In T. Husen & T. N. Postlethwaite (Eds.), *The international encyclopedia of education* (2nd ed., pp. 671–678). 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). Erlbaum.

Acknowledgments

Figure 1 was adapted from "Adaptive Measurement of Change: A Novel Method to Reduce Respondent Burden and Detect Significant Individual-Level Change in Patient-Reported Outcome Measures" by D. J. Weiss et al., 2021, *Archives of Physical Medicine and Rehabilitation*. Copyright 2021 by The American Congress of Rehabilitation Medicine. Adapted with permission.

Figure 2 is reprinted from *Measuring Intra-Individual Change at Two or More Occasions with Hypothesis Testing Methods* by C. Phadke, 2017, Unpublished doctoral dissertation, University of Minnesota, Twin Cities. Copyright 2017 by C. Phadke. Reprinted with permission.

Citation

Tai, M. H., Cooperman, A. W., DeWeese, J. N., & Weiss, D. J. (2023).
How do trait change patterns affect the performance of adaptive measurement of change?
Journal of Computerized Adaptive Testing, 10(3), 32–58. https://doi.org/10.7333/2307-1003032

Author Address

Ming Him Tai
Email: tai00006@umn.edu