Journal of Methods and Measurement in the Social Sciences
Vol. 2, No. 1, 80-101, 2011

Measuring Individual Growth With Conventional and Adaptive Tests

David J. Weiss and Shannon Von Minden
University of Minnesota

Measuring individuals or groups longitudinally is frequently necessary in social science research and applications. Substantial research and discussion has focused on the statistical properties of measures of change and some of the psychometric problems involved This monte-carlo simulation study focused on properties of the measurement instruments used for obtaining scores that represent change or growth over five time points and examined how well scores from conventional tests and computerized adaptive tests used to measure individual growth curves reflect true change. Data representing four different patterns of individual change and a baseline no-change condition were generated from an item response theory (IRT) model. Different tests simulated were conventional peaked tests with narrow and wider difficulties and three levels of discrimination, and computerized adaptive tests (CATs) drawn from banks with the same levels of discrimination. Conventional tests were scored by number correct and IRT weighted maximum likelihood. Results showed that as the examinees' scores moved from the difficulty levels at which the tests were concentrated, number-correct scores over-estimated true change and had increasing amounts of error. High discrimination conventional tests had the poorest recovery of change for both groups and individuals. IRT scoring of the conventional tests improved recovery of change somewhat. By contrast, CATs consistently estimated growth with minimum and consistent error and performed best with highly discriminating items.

Keywords: adaptive testing, computerized adaptive tests, conventional tests, individual growth, item response theory, measuring change, measuring growth, off-target tests

Frequently in social science research and applications, people are measured on more than a single occasion. In many cases, interest is in changes over time that occur at the group level. For example, a researcher might be interested in changes over time in attitudes or perceptions of different groups of people, such as those affiliated with different political parties. In other instances, a social science researcher might perform an experiment that measures a group on one occasion, applies some experimental procedure and/or treatment, and then measures the same group of individuals on another occasion. In medical research, following measurements on the depression of a group of patients, a treatment is prescribed and the measurement of depression is repeated at a later date or a series of later dates. In educational research, different teaching approaches might be used in different schools and classrooms and group gains from a baseline measurement are examined. Developmental

researchers frequently are concerned with patterns of group growth over time for different cognitive abilities.

Repeated measurements also frequently are obtained and examined for single individuals to evaluate patterns of growth or decline. In schools, it can be important to track how well a student is doing with regard to learning a defined body of knowledge across a period of weeks or a semester, with measurements taken at frequent intervals. Medical researchers might be interested in patterns of decline in cognitive functioning of the elderly across time for specific cognitive tasks. Similarly, those working with children with special needs might need to monitor their intellectual growth by measuring then monthly for a period of a year or more to determine if improvement is occurring.

Research on measuring change—including growth, decline, and lack of change—has primarily been focused on change over two occasions. Although Cronbach and Furby (1970) proposed a moratorium on attempting to measure change based on some psychometric issues that they identified that suggested that measures of change over two occasions were psychometrically flawed, because of the applied and research need to examine both group and individual change, attempts to resolve the psychometric issues have continued (e.g., Bereiter, 1963; Collins, 1996; Hummel-Rossi & Weinberg, 1975; Lord, 1963; Mellenbergh, 1999; Overall & Woodward, 1975; Rogosa & Willett, 1983; Willett, 1997; Williams & Zimmerman, 1996a, 1996b; Zimmerman & Williams, 1982), focusing almost entirely on change observed at two occasions. Change across two occasions has also been examined in the context of item response theory and computerized adaptive testing by May and Nicewander (1998) and Kim-Kang and Weiss (2008).

A common approach to measuring growth, whether at two occasions or across multiple occasions, has been the repeated use of the same measuring instrument. This frequently occurs in "pre-post" studies and applications, but also when individuals are measured across multiple occasions. The same instrument is used in many instances because alternate or "parallel" forms of many psychological measuring instruments are not available due to the expense involved in constructing parallel forms and the difficulty of creating two or more forms of a test that function equivalently. Also, in many cases, the measuring instruments used have been constructed according to classical test development procedures, which are designed to maximize internal consistency reliability by selecting items for the instrument that (1) have high discriminations and (2) have item difficulties (proportion correct) around .50, or for non-dichotomous items mean total scores that are at the center of the response scale range.

However, when the same or "parallel" forms of a given measurement instrument are used to measure growth (or decline) for either groups or

individuals, the researcher runs the risk of the measuring instrument becoming off-target (Embretson, 1996; Kang & Waller, 2005) for the examinees when growth or decline has occurred. Weiss (2011) briefly described the potentially detrimental effects of off-target measuring instruments on scores from conventional tests, and conclusions drawn from them, and proposed computerized adaptive testing (CAT) as a viable solution. This study further examined this problem in the context of measuring both group and individual change across multiple occasions.

Purpose

The purpose of this study was to examine how well scores from conventional tests and CATs (Weiss, 2011) used to measure individual growth curves reflect true change. A conventional test is composed of a fixed set of items and, typically if it is a "peaked" test, it is targeted at a specific range of the ability or trait being measured, which might or might not correspond to a given examinee's true trait level when change has occurred. The current simulation study was an examination of the similarity between true growth curves and observed growth curves across five measurement occasions, for both groups and individuals. Factors that might impact how well a test reflects true change include where and how much the test is peaked, the item discriminations, the way the test is scored, and how the test is administered (conventionally or adaptively).

Method

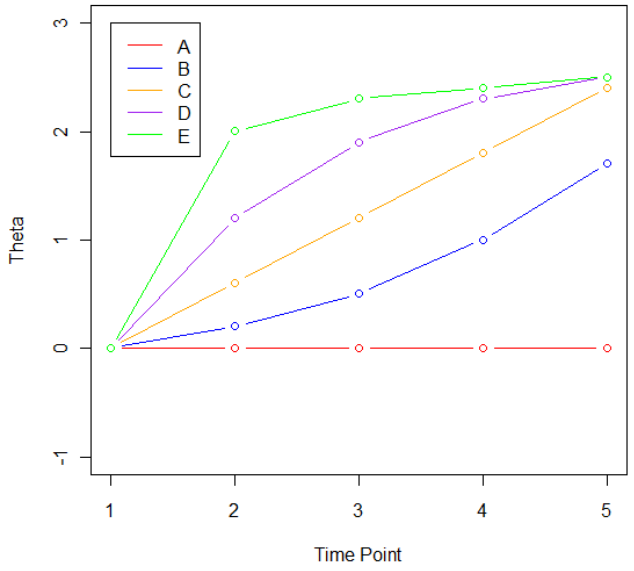
True θ Values

Using monte-carlo simulation, the Time 1 true ability/trait (θ) values (θ_1) for 200 simulated examinees (simulees) were drawn from a uniform distribution ranging from −.25 to .25. Five different true growth curves were simulated using the same 200 simulees in each condition. Growth Curve A was used as a baseline in which the simulees' true θ levels did not change over the five time points. The other four growth curves were generated by adding a constant to the θ_1 values for each of the simulees at each of the four additional time points. Table 1 shows how the growth curves were simulated and Figure 1 illustrates the mean growth curves. Growth Curve B reflected slow but accelerating growth, Curve C was linear growth, and Curves D and E reflected decelerating growth, with higher initial growth for Curve E.

Table 1
The Five Growth Curves

Curve	Time 1	Time 2	Time 3	Time 4	Time 5
A	θ_1	θ_1	θ_1	θ_1	θ_1
B	θ_1	$\theta_1 + .2$	$\theta_1 + .5$	$\theta_1 + 1$	$\theta_1 + 1.7$
C	θ_1	$\theta_1 + .6$	$\theta_1 + 1.2$	$\theta_1 + 1.8$	$\theta_1 + 2.4$
D	θ_1	$\theta_1 + 1.2$	$\theta_1 + 1.9$	$\theta_1 + 2.3$	$\theta_1 + 2.5$
E	θ_1	$\theta_1 + 2$	$\theta_1 + 2.3$	$\theta_1 + 2.4$	$\theta_1 + 2.5$

Figure 1. Mean True Growth Curves



Conventional Tests

Scored item responses to 50-item conventional tests were generated using the 3-parameter logistic IRT model

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

where P_{ij} is the probability of a correct response to item i by examinee j, θ_j is the trait level of examinee j, a_i is the discrimination of item i, b_i is the difficulty of item i, c_i is the pseudo-guessing parameter of item i, and D = 1.7. Three item discrimination conditions were generated. Each condition had a normal distribution with a mean a_i = .6, 1.0, or 1.5, representing low

(LD), medium (MD), and high (HD) discriminations, respectively. All of the item discrimination conditions had a standard deviation of .1. There were two item difficulty conditions, as well. Items for conventional tests are commonly selected to have difficulties close to b_i = 0.0 in order to maximize the item variance, which increases the internal consistency reliability of the test (Crocker & Algina, 2006). Thus, the item difficulties for the conventional tests were simulated to be closely centered around b_i =0.0. In the Narrow *b* condition, b_is were generated from a uniform distribution ranging from b_i = −.5 to .5. In the Wide *b* condition, b_is were generated from a uniform distribution ranging from b_i = −1 to 1. The c_i parameter was set to .2 for all items. Randomly parallel 50-item tests were generated at each time point.

Item responses were simulated using the true θ values for examinees at each time point using a program written in R (R Development Core Team, 2010). A matrix containing the expected probabilities of correct responses was calculated using Equation 1. A matrix of random numbers was also generated from a uniform distribution ranging from 0 to 1. If the random number generated for a simulee on a given item was less than the expected probability of a correct response for that item, the item was scored as correct (1). If the random number was greater than or equal to the expected probability, the item was scored as incorrect (0).

The number-correct scores were then calculated for each time point. Using the test response function associated with each set of item parameters for each set of 50 items, with θ values incremented by .05, number-correct scores were transformed to the θ metric (N-C θ). This allowed the observed θ approximations and the true θ values to be compared on the same scale. IRT weighted maximum likelihood (WML; Guyer, 2010, p.37; Warm, 1989) θ estimation was also used to estimate θ levels to investigate whether scoring the same data by IRT improved the recovery of the true growth curves. WML was used rather than maximum likelihood because it can provide θ estimates for all correct or all incorrect response patterns.

Computerized Adaptive Tests

Scored item responses to item banks used to administer 50-item fixed-length CATs were generated using Equation 1. The discrimination conditions (low, medium, and high) for the CATs were the same as those used in the conventional tests, and c_i = .20 for all items, as well. The range of item difficulties for each CAT item bank was between b_i = −3.50 and 3.50. The difficulty range was broken into 14 segments, each of which had a width of .50. Each segment contained 25 items, which totaled to 350 items in each item bank. Within each segment, the b_is were generated

from a uniform distribution. There were three CAT item banks used—one for each discrimination condition.

Responses to all of the items in the CAT item banks were simulated using R (R Development Core Team, 2010) using the same method described above. CATSim (Weiss & Guyer, 2012) was then used to simulate the administration of a CAT for each time point. Initial θ estimates at Time 1 were set to 0.0 for all simulees. For time points 2 – 5, the initial θ estimate for each simulee was the final θ estimate from the previous time point, as proposed by Weiss and Kingsbury (1984). The θ estimates were obtained using maximum likelihood with a step size of 3 for non-mixed response patterns. Items were selected using maximum information, selecting at each stage in the CAT the unadministered item in the 350-item bank that provided maximum information at the current θ estimate. All of the CATs were terminated after 50 items were administered.

Analysis

In order to evaluate how close an observed point on the mean growth curve was to the corresponding true growth curve, RMSE was calculated. RMSE averaged across simulees for occasion *k* is given by

$$RMSE_k = \sqrt{\frac{1}{N}\sum_{j=1}^{N} \hat{\theta}_{kj} - \theta_{kj}^{\ 2}} \qquad (2)$$

where $N = 200$, $\hat{\theta}_{kj}$ is the estimated θ value of simulee *j* at time *k*, and θ_{kj} is the true θ of simulee *j* at time *k*.

Bias was calculated to evaluate whether change was being overestimated or underestimated at each time point. Bias is given by

$$Bias_k = \frac{1}{N}\sum_{j=1}^{N} \hat{\theta}_{kj} - \theta_{kj} \qquad (3)$$

Positive bias indicates that true change was being overestimated and negative bias indicates that true change was being underestimated.

In addition to calculating the RMSE at each time point to see how it changed as θ moved farther from the targeted ability level, RMSE was examined for each simulee's growth curve to see how well individual growth was recovered. The RMSE averaged across time points for simulee *j* given by

$$RMSE_j = \sqrt{\frac{1}{5}\sum_{k=1}^{5} \hat{\theta}_{kj} - \theta_{kj}^{\ 2}}. \qquad (4)$$

Results

Group Growth

Means and SDs

Table 2 shows the means and standard deviations (SD) of true θ at each of the five time points, as well as N-C θ, WML θ, and CAT θ for the six combinations of *a* and *b* for the baseline Curve A (no-change) condition. Under all item bank conditions, CAT had a tendency to slightly overestimate true θ and conventional tests had a tendency to slightly underestimate true θ, whether scored by N-C or WML. The SDs for all θ estimation methods were higher than the SDs of true θs, reflecting error of measurement. SDs were highest for the Low *a* condition and decreased as *a* increased, for both the Narrow and Wide *b* conditions. For the Narrow *b* conditions, CAT θ estimates generally had lower SDs than the conventional tests, whether scored by N-C or WML. For the Wide *b* conditions, CAT θ estimates had smaller SDs for all but two conditions.

Figure 2 shows the same data as in Table 2, but for Curve B, a slowly accelerating growth curve (numerical values for all figures are in Von Minden, 2011). For both the Narrow *b* (Figures 2a – 2c) and Wide *b* (Figures 2d – 2f) conditions, mean estimated θ s were close to mean true θs for Times 1 – 4. At Time 5, however, when items had high *a*s, the true mean θ was 1.677 but the NC mean θ was 2.200 for the Narrow *b* condition and 1.883 for the Wide *b* condition.

Figure 2 also displays differential effects on the SDs of the θ estimates at different time points, as indicated by the vertical lines plotted at the means for each condition. True θ SDs were .142 for all time points. For the Narrow *b* condition (Figures 2a – 2c), SDs of CATs θs were essentially constant across time points for a given level of *a*: for Narrow *b*, they ranged from .326 to .374 for Low *a,* .254 to .285 for Medium *a,* and .195 to .227 for High *a*. By contrast, SDs of N-C θs increased as the mean true θ increased, ranging from .372 to .519 for Low *a* (Figure 2a), .293 to .484 for Medium *a* (Figure 2b), and .229 to .834 for High *a* (Figure 2c), the latter result indicating substantial amounts of error in these θ estimates. WML SDs increased with increasing means for the Low and Medium *a* conditions, but remained relatively constant for the High *a* condition.

A similar pattern of results for the SDs was observed for the Wide *b* conditions, but the differences were less pronounced. CAT θ SDs were low and essentially constant across time points within an item discrimination level, whereas the SDs of the N-C θs increased as the means increased; the maximum NC SDs were all at Time 5 with values of .502, .467, and .563 for Low, Middle, and High *a*, respectively, versus .142 for true θs.

Table 2

Means and SDs of θ Estimates for Each Condition at Each Time Point, for Curve A

	Time 1		Time 2		Time 3		Time 4		Time 5	
Condition	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
True θ	−.023	.142	−.023	.142	−.023	.142	−.023	.142	−.023	.142
Narrow *b*, Low *a*										
N-C θ	−.028	.372	−.026	.373	−.013	.405	.022	.391	−.019	.400
WML θ	−.029	.362	−.034	.346	−.020	.390	.015	.380	−.014	.383
CAT θ	.030	.374	.019	.353	.042	.377	.033	.375	.007	.329
Narrow *b*, Medium *a*										
N-C θ	−.003	.293	−.007	.281	−.017	.263	−.013	.263	−.020	.286
WML θ	−.005	.276	−.008	.272	−.015	.251	−.006	.248	−.020	.274
CAT θ	.033	.260	−.011	.266	−.010	.271	.020	.263	.010	.278
Narrow *b*, High *a*										
N-C θ	−.012	.229	.015	.203	−.009	.208	−.016	.214	−.013	.201
WML θ	−.008	.214	.007	.186	−.005	.204	−.017	.213	−.016	.194
CAT θ	.014	.227	−.013	.208	.011	.210	−.004	.230	−.016	.222
Wide *b*, Low *a*										
N-C θ	−.035	.393	−.026	.391	.015	.414	−.008	.392	−.015	.385
WML θ	−.037	.372	−.004	.376	.011	.402	−.016	.376	−.023	.375
CAT θ	.030	.374	.019	.353	.042	.377	.033	.375	.007	.329
Wide *b*, Medium *a*										
N-C θ	−.056	.271	−.008	.303	−.038	.282	−.048	.274	−.003	.253
WML θ	−.062	.258	−.011	.289	−.029	.268	−.031	.257	−.003	.253
CAT θ	.033	.260	−.011	.266	−.010	.271	.020	.263	.010	.278
Wide *b*, High *a*										
N-C θ	−.007	.229	−.044	.215	−.026	.235	−.017	.219	−.047	.255
WML θ	−.008	.219	−.041	.208	−.026	.227	−.014	.214	−.045	.238
CAT θ	.014	.227	−.013	.208	.011	.210	−.004	.230	−.016	.222

Figure 2. Means and SDs (Vertical Lines) of θ Estimates at Five Time Points for Growth Curve B

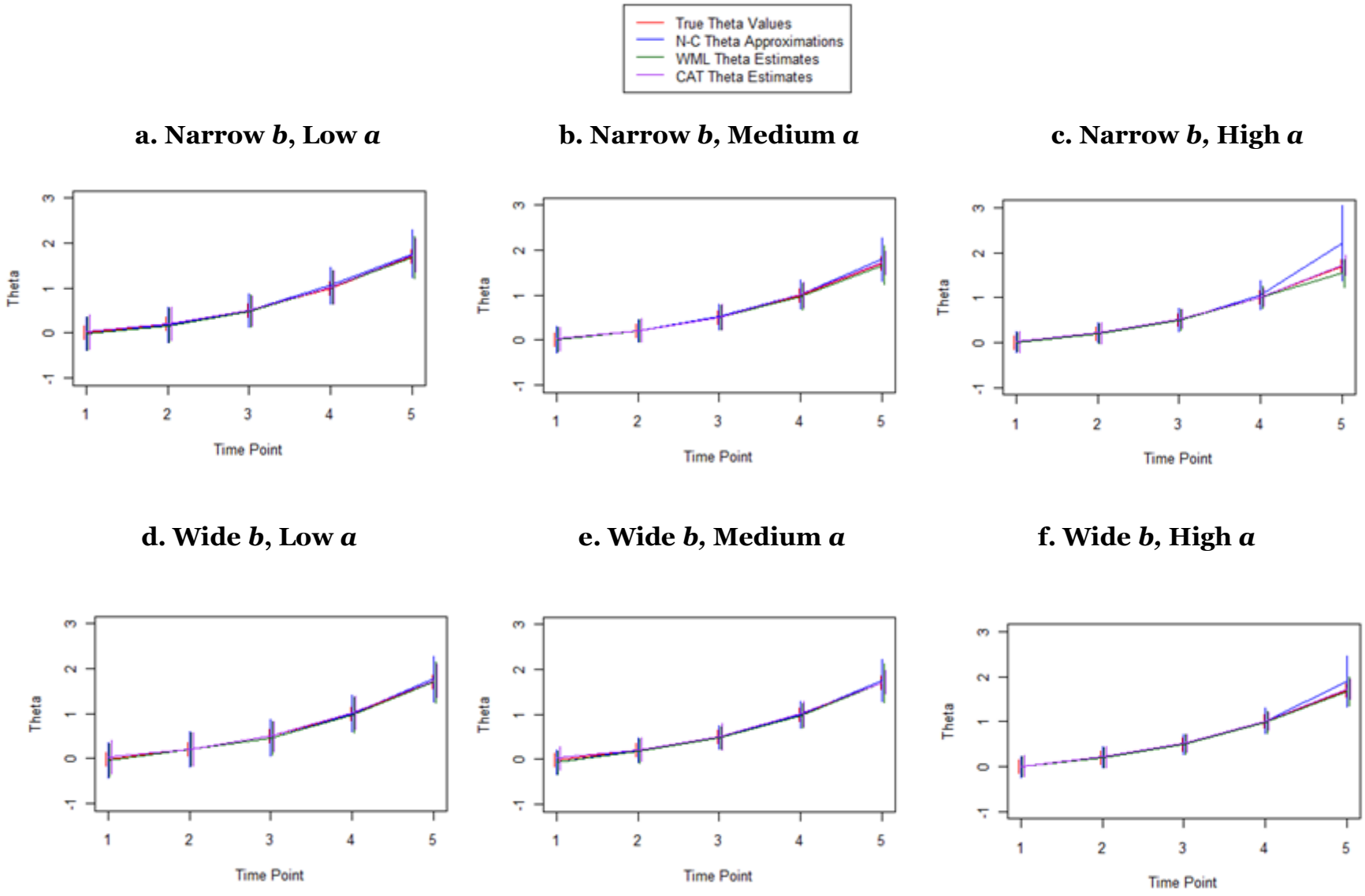


Figure 3 shows means and SDs of estimated θ for Curve C, which modeled linear growth. Similar results were obtained as were observed for Curve B—as the mean true θ deviated further from 0.0, where the conventional test was peaked, mean θ estimates for the conventional test (both N-C and WML) deviated further from mean true θ. However, mean NC θ estimates tended to be higher than mean true θ, whereas mean WML estimates (based on the same item responses from which N-C scores were computed), tended to be lower than the true means (e.g. Figures 3c and 3f), with a somewhat diminished effect for the Wide *b* tests. Mean θs for CATs were all very close to the true means for all time points and all combinations of *a* and *b* conditions.

As for Curve B, SDs of N-C θs increased over time points for all *a* and *b* conditions. Maximum N-C SDs with Narrow *b* were .525 at Time 4 for Low *a*, .494 for Medium *a,* and .760 for High *a* (Time 5 SDs were lower due to ceiling effects). Wide *b* tests had little effect on the SDs—corresponding SDs were .530, .460, and .777, respectively. By contrast, CAT θ SDs were constant across time points within an *a* level, for both Narrow and Wide *b* tests.

The means and SDs of the true and estimated θ values for Curve D, a decelerating growth curve, are shown in Figure 4. The simulees' true θ values were within the targeted range of the conventional test only at Time 1. After that time point, using the Narrow *b* test (Figures 4a – 4c) the N-C θ approximations (M = 1.311, 2.454, 2.759, and 2.874) increasingly overestimated the true means (M = 1.177, 1.877, 2.277, and 2.477) and had the largest standard deviations (SD = .603, .740, .520, and .390) in the High *a* condition. The WML θ estimates (M = 1.124, 1.665, 1.796, and 1.749) underestimated the true means and had the smallest standard deviations (SD = .291, .261, .170, and .120) in the High *a* condition. As before, the mean CAT θ estimates (M = 1.207, 1.910, 2.306, and 2.503) were very close to the true means, and the standard deviations (SD = .224, .210, .212, and .217) were smallest in the High *a* condition.

A similar pattern of results was observed for the Wide *b* conditions (Figures 4d – 4f), although deviations from true values were somewhat attenuated. N-C θ s were higher than true θs, particularly in the High *a* condition, where the magnitude of overestimation tended to increase with increasing distance of the true means from the location of the test. SDs of N-C θs increased at successive time points for all *a* conditions, reaching their maximum of .717 at Time 3 for the High *a* condition, before reducing due to ceiling effects. Again, WML mean θs were close to the true means for the first three time points, then underestimated true θs. SDs of WML θs were small and constant across time points, and decreased with increasing *a*. CAT mean θs were again very similar to true θs and, similar to WML SDs, CAT SDs remained constant across time points, were generally the smallest of the θ estimation methods, and decreased with increasing *a*.

The means and SDs for the true and estimated θ values for Curve E are shown in Figure 5. As with Curve D, the simulees' true θs fell within the targeted range of the conventional test only at Time 1. After that, the N-C θs again overestimated the true mean θs and the WML θs underestimated the true means, with the most pronounced

Figure 3. Means and SDs (Vertical Lines) of θ Estimates at Five Time Points for Growth Curve C

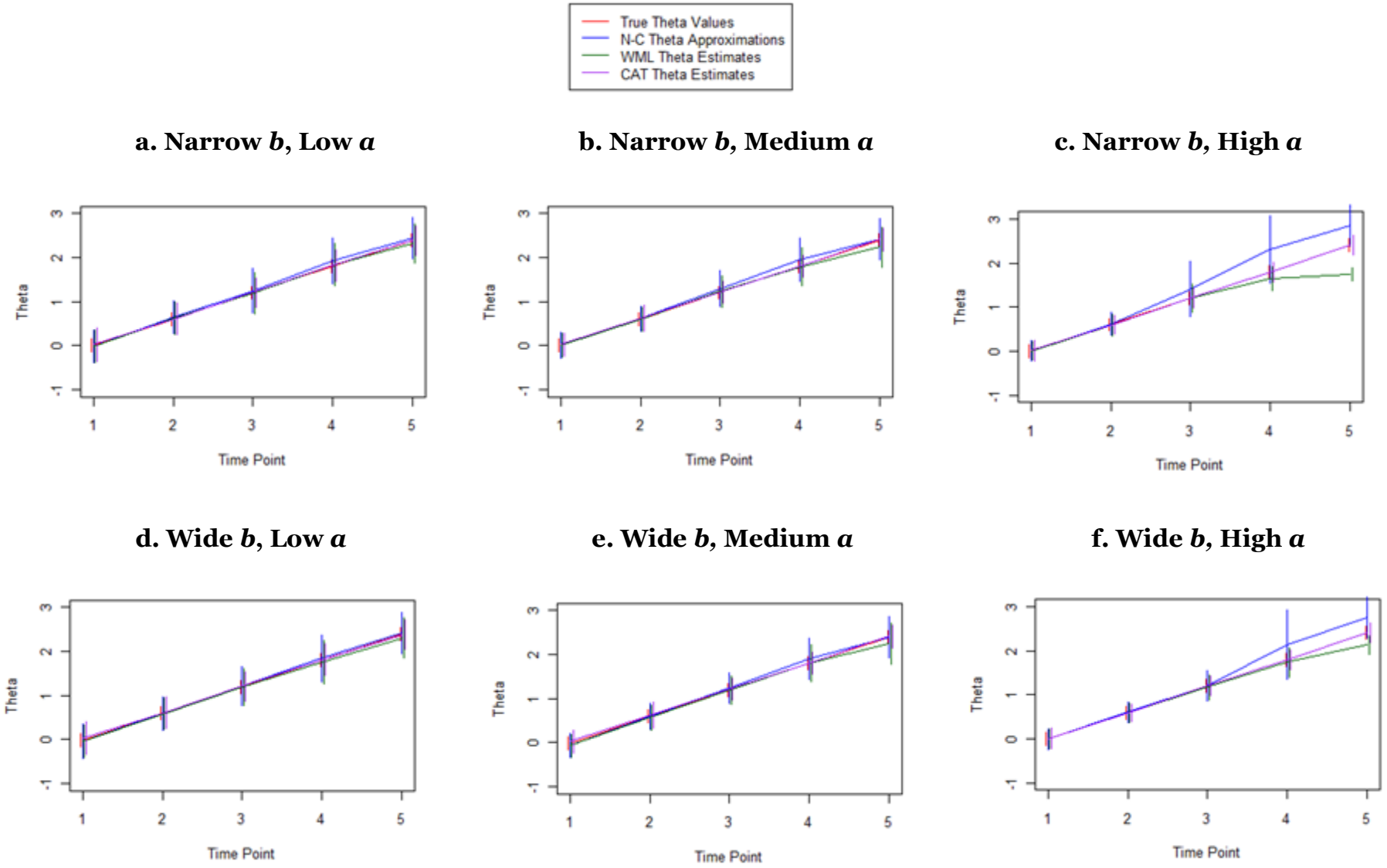


Figure 4. Means and SDs (Vertical Lines) of θ Estimates at Five Time Points for Growth Curve D

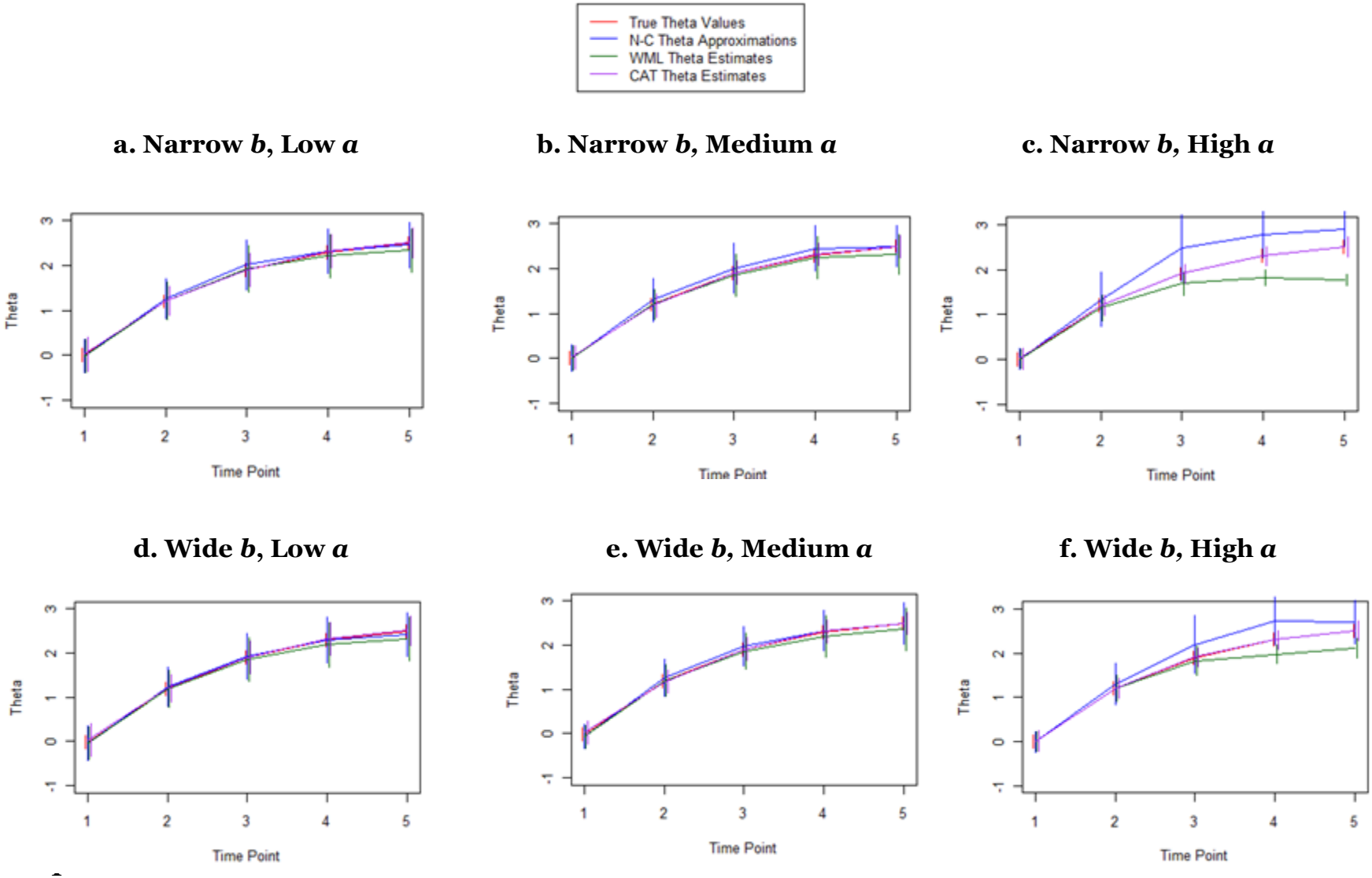
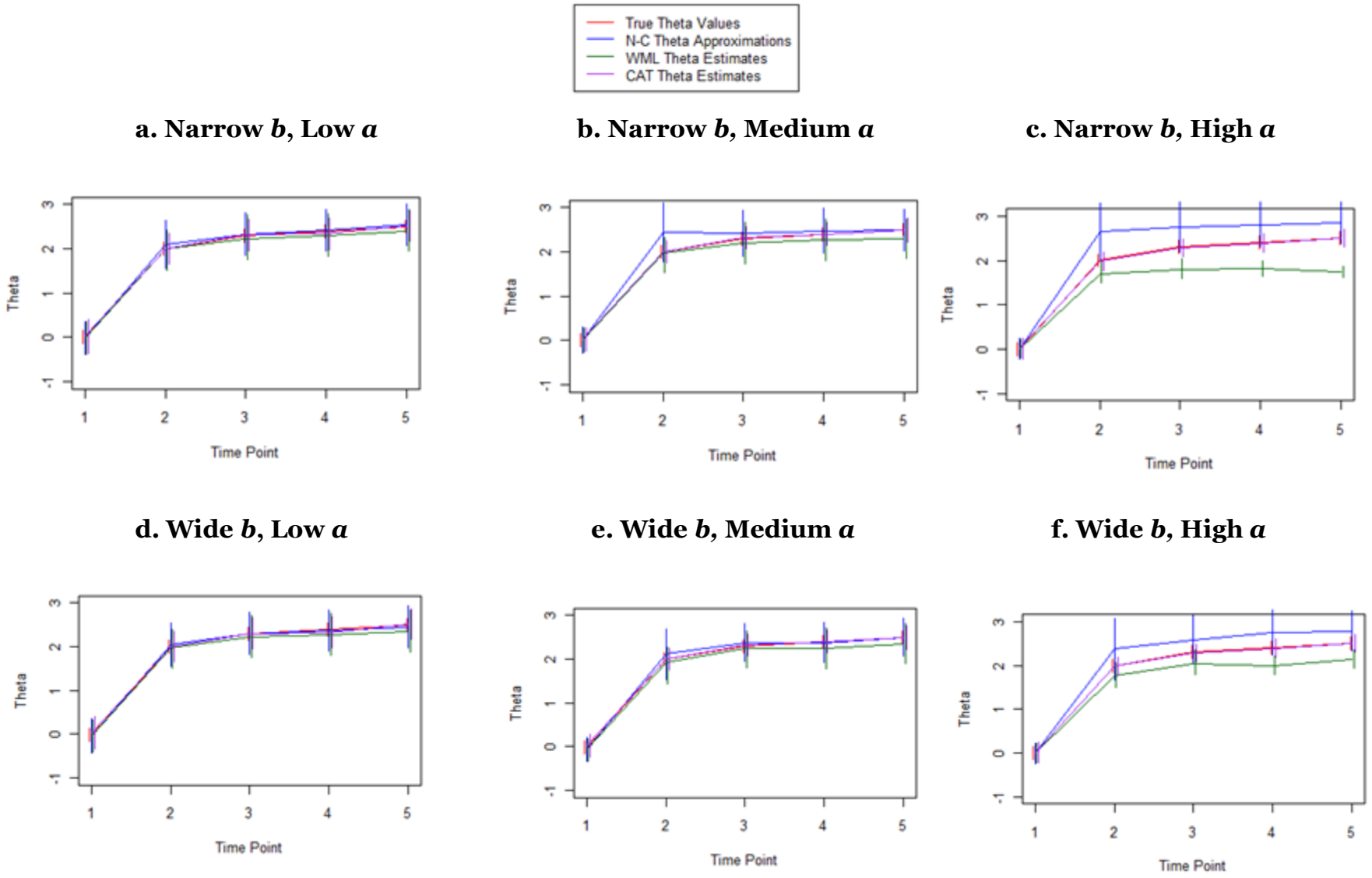


Figure 5. Means and SDs (Vertical Lines) of θ Estimates at Five Time Points for Growth Curve E



differences in the High *a* conditions. For Time 2 – 5 for the Narrow *b* condition (Figures 5a – 5c), true means were 1.977, 2.277, 2.377, and 2.477; by contrast, N-C means were 2.637, 2.721, 2.784, and 2.821; and WML means were 1.677, 1.761, 1.804, and 1.733. As with other growth curves, the mean CAT θ estimates for Times 2 – 5 (M = 1.983, 2.279, 2.392, and 2.504) were close to the true means.

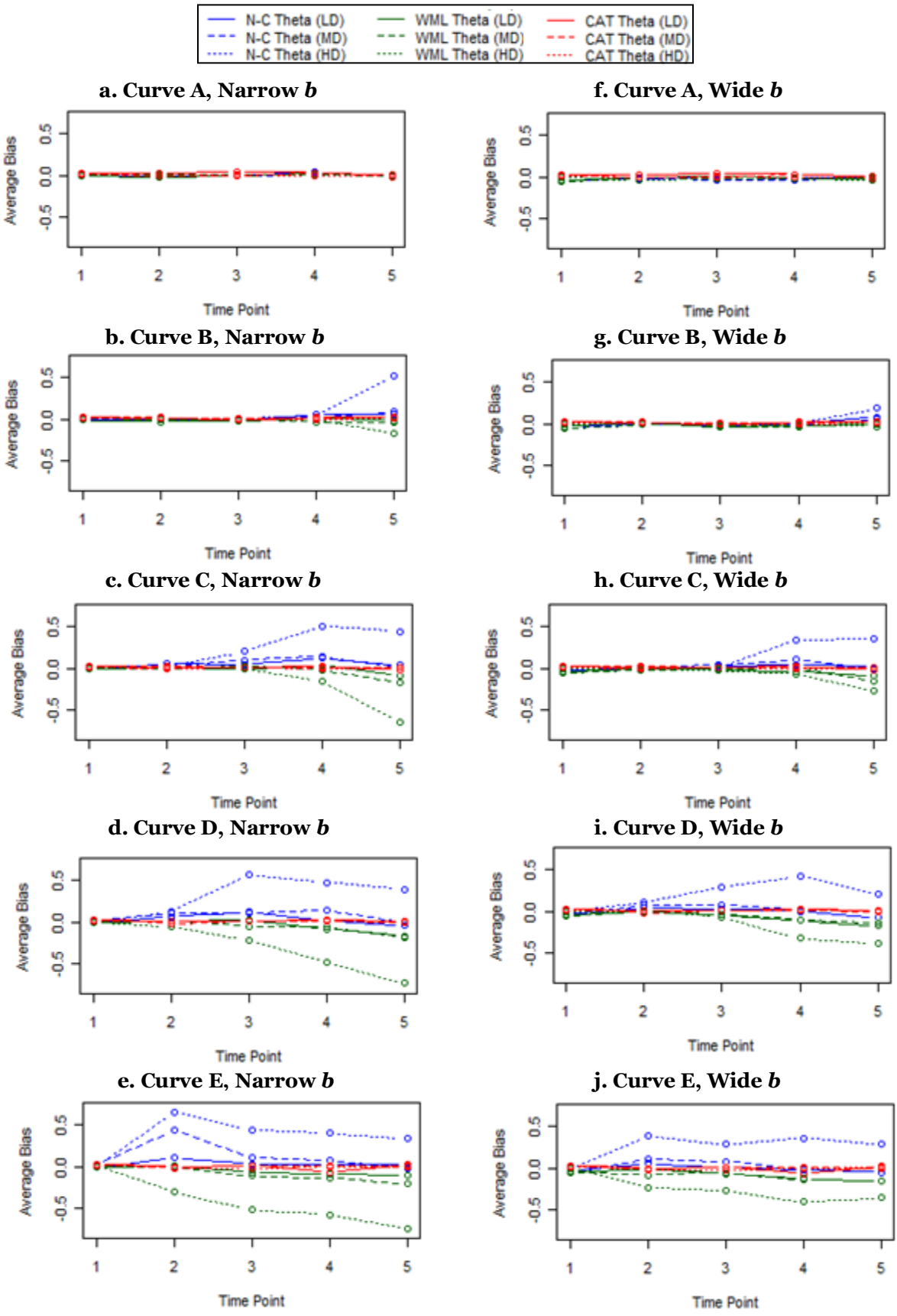
The SDs followed the same pattern as the previous growth curves, as well; in the High *a* condition they were largest for the N-C θ approximations and smallest for the WML and CAT θ estimates. NC θ SDs were as high as .668 for the Narrow *b* tests and .700 for the Wide *b* tests, compared to .142 for true θ, approximately .230 for CAT θ, and about .250 for WML θ with High *a* items.

Bias

The average bias at each time point for all of the growth curves and θ estimation methods is shown in Figure 6. As was seen with the mean θ estimates, N-C θ estimates tended to be positively biased when the mean true θ differed from the test difficulty, the WML θ estimates tended to be negatively biased, and the CAT θ estimates were essentially unbiased. The magnitude of the bias was not always larger for one type of θ estimation under the Narrow *b* condition (Figures 6a – 6e). At some time points, the N-C θ approximations had noticeably larger values of bias, such as Time 5 in Curve B (Figure 6b) or Time 4 (Figure 6c) in Curve C. However, there were time points where the WML θ estimates had noticeably larger values of bias, such as Time 5 in Curves C, D, and E (Figures 6c – 6e) for the Narrow *b* conditions. As the difference between the targeted θ range of the conventional tests and the true θ values of the simulees became larger, the average bias for the WML θ estimates became larger in magnitude than the N-C θs, although in the opposite direction. For N-C θ and WML, the bias was largest in magnitude in the High *a* (HD) condition and at time points at which the simulees had moved farther away from the conventional tests.

For the Wide *b* condition, a similar picture emerged (Figures 6f through 6j). Although CAT θs remained unbiased, as they had for the Narrow *b* conditions, bias for the WML θs was reduced so that WML had slightly larger bias than N-C θs for High *a* (HD) conditions only at Time 4 for Curve D (Figure 6i) and Times 4 and 5 for Curve E (Figure 6j).

Figure 6. Average Bias for Each Growth Curve for Narrow *b* and Wide *b* Conditions



RMSE

RMSE, the SD of the difference between estimated θ and true θ across a group of simulees, is a direct indicator of the error of measurement for a given θ estimation method and condition The average RMSEs at each time point for all of the growth curves are shown in Figure 7. For the time points at which the simulees' true θs were within the targeted range of the conventional tests, the High *a* condition had the lowest average RMSE and the Low *a* condition had the highest average RMSE. Once the simulees moved far outside the targeted range of the conventional tests, however, the High *a* condition had the highest average RMSE (for N-C θ and WML θ). Generally, N-C θ had higher average RMSEs than WML θ; the CAT θ estimates had the lowest average RMSE values across all discrimination conditions, with RMSE decreasing with increasing *a* and essentially constant across time points regardless of the distance between true θ and the location of the test.

The magnitudes of RMSE were quite large under several conditions for the N-C θs. For Curve B with Narrow *b* at Time 5, RMSE approached a full θ SD unit (Figure 7b). RMSEs of N-C θs approached or exceeded .75 SD units for at least one time point for Curves C, D, and E for both the Narrow *b* condition (Figures 7c – 7e) and the Wide *b* condition (Figures 7h -7j). RMSEs were near .75 for WML only for Curves D and E at Time 5 for Narrow *b* conditions.

Recovery of Individual Growth Curves

The RMSE for individual examinees reflects the SD of the differences between the estimated growth curve for a simulee and the simulee's true growth curve. A smaller RMSE indicates better recovery of the true growth curve. The distributions of RMSE for individual simulees across all five time points, conditions, and θ estimation methods are shown in Figure 8.

Regardless of growth curve, CAT θ had RMSE values that were the same or smaller than those for the N-C θ or WML θ. CAT θs had the smallest RMSE values in the High *a* condition, followed by the Medium then Low *a* conditions. For Curve E in the Narrow *b* condition (Figure 8e), CAT θs had a mean RMSE of .143 with SD = .046 and a range of .048 to .260; for the same conditions, NC θ RMSEs had M = .642, SD = .114, and ranged from .444 to .888. For the same condition with a Wide *b* test, there was also no overlap between the CAT and N-C θ RMSE distributions.

The distributions of RMSE values for the N-C θs and WML θs changed depending on the growth curve. For Curve A (no growth), the High *a* condition with Narrow *b* (Figure 8a) clearly measured individuals with the most precision because it had the lowest individual RMSEs for all θ estimation methods; that trend was not obvious when the tests had Wide *b* (Figure 8f).

Figure 7. Average RMSE for Each Growth Curve for Narrow *b* and Wide *b* Conditions

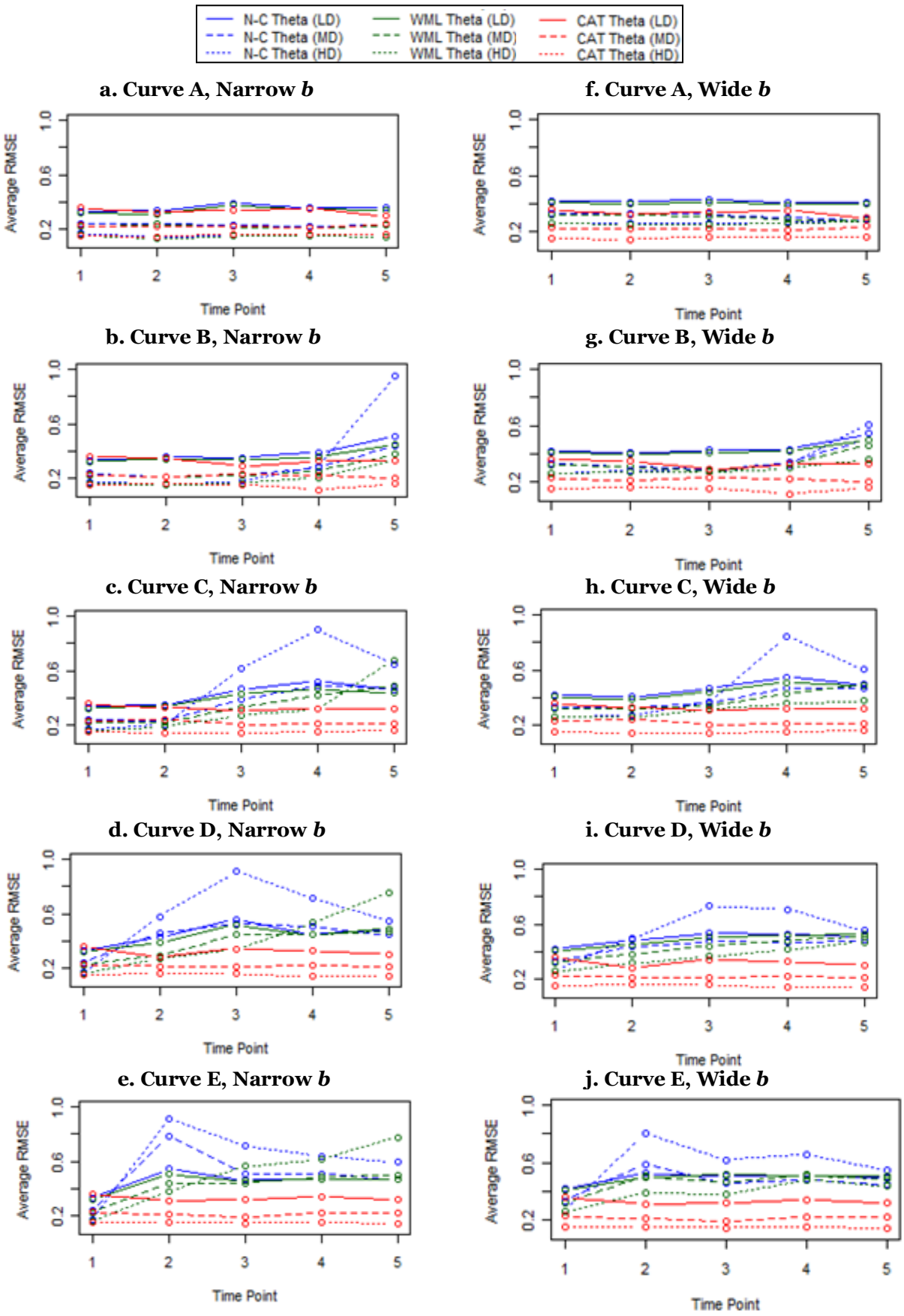
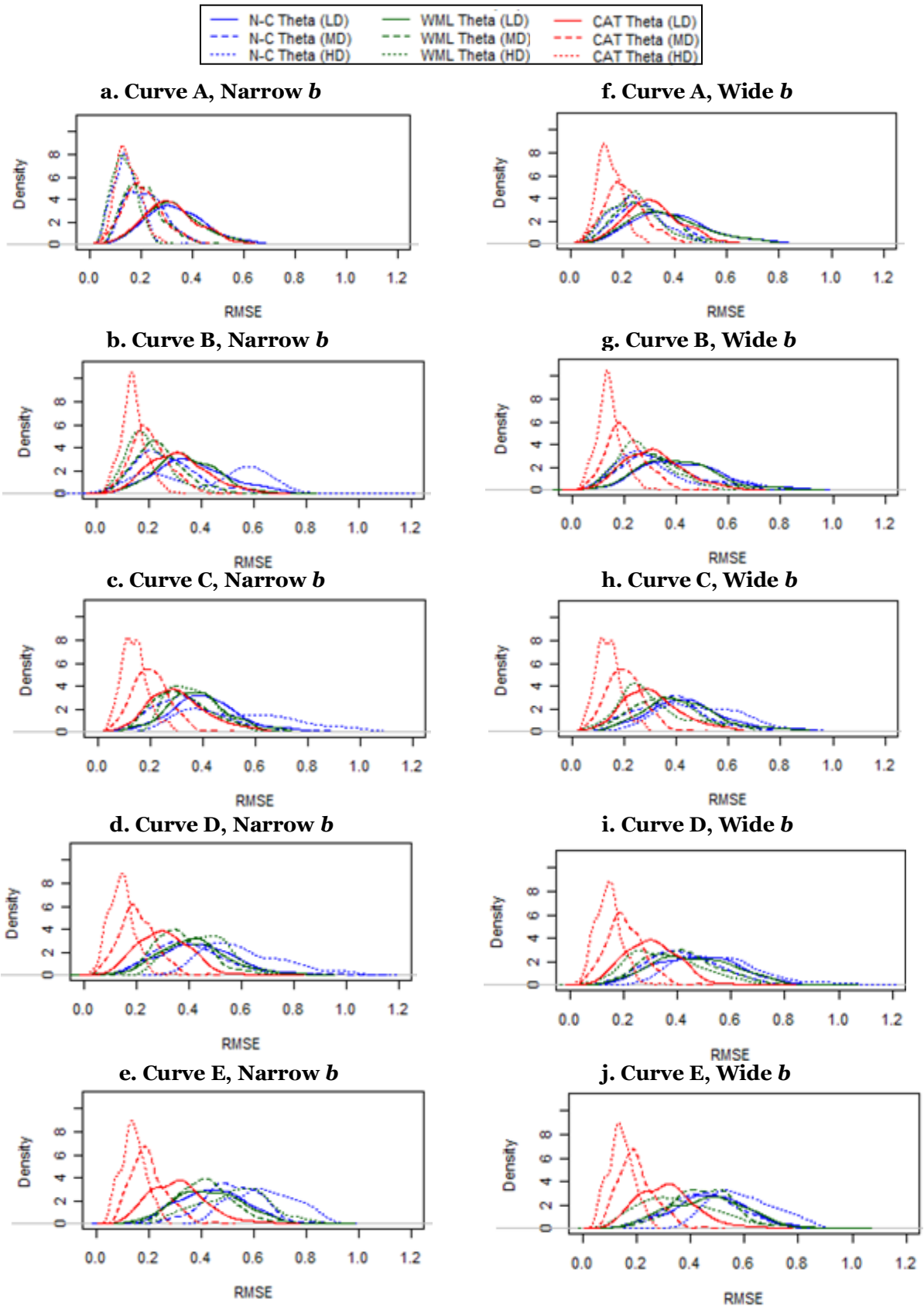


Figure 8. Distribution of Individual RMSEs for Each Growth Curve for Narrow *b* and Wide *b* Conditions



For Curves B through E with Narrow *b*, the RMSEs in the Low and Medium *a* conditions did not change dramatically, but the RMSEs for the N-C θs in the High *a* condition became considerably larger. Again, this trend was somewhat less obvious for the Wide *b* condition. Across all of the growth curves, the WML θ estimates had lower individual RMSEs than the respective N-C θ approximations, even though they were based on the same sets of item responses.

Discussion and Conclusions

Overall, tests consisting of highly discriminating items performed better (lower bias and RMSE in the recovery of true growth) than tests containing items with lower discriminations when the test's difficulty corresponded to the examinees' θ levels. An item with high discrimination has an information function that is peaked; the item provides substantial information at the particular θ level that corresponds to the item difficulty, but that high level of information is concentrated over a narrow range of θ. An item with low discrimination has an information function that is less peaked; it provides less information at the particular θ level that corresponds to the item difficulty, but that information is spread out over a wider range of θ. When items are aggregated into tests, tests comprised of highly discriminating items that are similar in difficulty, have highly peaked information functions, whereas tests with less discriminating items have test information functions that are lower and less peaked.

When the examinees' θ levels were outside the targeted range of the test (i.e., change had occurred), the high discrimination condition performed the worst when using conventional tests. This can be explained by the low level of information that is available at θ levels that do not correspond very closely to the item difficulties when using items with high discriminations. Items with lower discriminations do not provide high levels of information at a particular level of θ, but they do offer more information at extreme θ levels than highly discriminating items. When using CATs instead of conventional tests, highly discriminating items performed the best regardless of examinee θ levels.

Widening the targeted θ range of the conventional tests improved the measurement of individual change somewhat, but once examinees moved outside the targeted θ range, the same pattern of results emerged, although the magnitudes were smaller than those in the narrow difficulty condition. The higher individual RMSE values for Curve A (the no-change condition) in the wide difficulty condition can be explained by the fact that the tests with wider difficulty ranges had items that were more spread out over the θ range than tests in the narrow difficulty condition. The narrow tests had more items concentrated around the specific range of θ in which the examinees were located at all five time points, which resulted in more information available for those examinees. The tests with a wider difficulty range resulted in less information for the examinees in growth curve A, but more information for examinees in the other growth curves in which change occurred. Thus, compared to the narrowly peaked conventional tests, the wider conventional tests did a somewhat better job of

measuring individuals who changed over the five time points, but did not do as well for the individuals who did not change.

Conclusions

This study demonstrated that conventionally constructed and scored peaked tests cannot adequately measure individual growth at multiple time points—number-correct scores from peaked conventional tests over-estimated true trait levels, with greater overestimation occurring as the examinee's trait levels moved further from the range of the test. IRT scoring of the conventional tests improved the recovery of true growth curves, but there was still a wide range of error in recovery of patterns of individual growth. By contrast, CATs recovered change very well, regardless of the pattern or level of change. These results support and extend the conclusions drawn by May and Nicewander (1998) and Kim-Kang and Weiss (2008), who independently concluded that change scores across two occasions from conventional tests poorly measure actual individual change. Both studies also showed that IRT scoring of conventional tests resulted in better recovery of true change, but also that change scores from adaptive tests recovered true change considerably better than either number-correct or IRT-scored conventional tests.

Researchers using the same test or parallel conventional tests to measure individual change over multiple time points have no way of knowing whether the observed growth curves reflect true change or measurement error, as demonstrated by the results of recovering individual patterns of growth. Without knowing how much an examinee has changed from one time point to another, it is impossible to know where to target the difficulty of the tests in order to improve the measurement of individual change. Tests designed according to the rules of classical test theory—highly discriminating items that are peaked or concentrated in difficulty—measure change more inaccurately than tests that deviate from these objectives. In other words, highly "reliable" conventional tests are poor tests for measuring individual growth or decline.

CAT provides a solution to the problem of not knowing how much each examinee changed over multiple testing occasions—and consequently, where to target the test's difficulty. Due to the individualized and dynamic item selection process for each examinee, using CAT to measure individual change allows for precise measurement of both examinees who do not change from one time point to another, as well as examinees who change dramatically. This study used fixed-length CATs; however, similar results could be expected using variable-length CATs (e.g., Finkelman, Weiss, & Kim-Kang, 2010), with consequent savings in numbers of items administered at each time point.

The growth curves used in this study might have been more realistic if there had been more individual variability in the pattern of growth. Regardless of that fact, the results of this study would generalize to any pattern of growth (or decline) in which an examinee's ability or trait level is outside the concentration of difficulty of a particular set of items. The results also generalize to tests that do not have sufficient numbers of items near an examinee's trait level as it changes on repeated measurements. Only adaptive tests operating from an item bank of

wide-ranging item difficulties can accurately measure individual patterns of growth or decline over repeated measurements—conventionally constructed tests cannot accurately measure change.

Corresponding author: David J. Weiss, Department of Psychology, University of Minnesota, N660 Elliott Hall, Minneapolis MN 55455-0344,
email: djweiss@umn.edu

References

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.

Collins, L. M. (1996). Is reliability obsolete? A commentary on "Are simple gain scores obsolete?" *Applied Psychological Measurement, 20,* 289-292.

Crocker, L., & Algina, J. (2006). *Introduction to classical & modern test theory*. Mason, OH: Thomson Wadsworth.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change" – or should we? *Psychological Bulletin, 74,* 68-8 .

Embretson, S. E. (1996). Item response theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement, 20,* 201-212.

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

Guyer, R. D. (2010). *Manual for ScoreAll 4.0: IRT Scoring for Conventionally Administered Tests*. St. Paul MN: Assessment Systems Corporation.

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).

Kang, S.-M. & Waller, N. G. (2005). Moderated multiple regression, spurious interaction effects, and IRT. *Applied Psychological Measurement, 29,* 87-105.

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

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

May, K. & Nicewander, W. A. (1998). Measuring change conventionally and adaptively. *Educational and Psychological Measurement,* 1998, *58,* 882.

Mellenbergh, G. J. (1999). A note on simple gain score precision. *Applied Psychological Measurement, 23,* 87-89.

Overall, J. E., & Woodward, J. A. (1975). Unreliability of difference scores: A paradox for measurement of change. *Psychological Bulletin, 82,* 85-86.

R Development Core Team. (2010). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Rogosa, D. B., & Willett, J. B. (1983). Demonstrating the reliability of the difference score in the measurement of change. *Journal of Educational Measurement, 20,* 335-343.

Von Minden, S. (2011). *Measuring individual change: A comparison of conventional and adaptive tests*. Unpublished Master's thesis, Department of Psychology, University of Minnesota

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54,* 427-45 .

Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences, 2(1)*, 1-23.

Weiss, D. J. & Guyer, R. (2012). *Manual for CATSim: Comprehensive simulation of computerized adaptive testing*. 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. (1997). Measuring change: What individual growth modeling buys you. In E. Arnsel & K. A. Reninger (Eds.), *Change and development* (pp. 213-243). Maywah, NJ: Erlbaum.

Williams, R. H., & Zimmerman, D. W. (1996a). Are simple gain scores obsolete? *Applied Psychological Measurement, 20,* 59-69.

Williams, R. H., & Zimmerman, D. W. (1996b). Commentary on the commentaries of Collins and Humphreys. *Applied Psychological Measurement, 20,* 295-297.

Zimmerman, D. W., & Williams, R. H. (1982). Gain scores in research can be highly reliable. *Journal of Educational Measurement, 19,* 149-154.