Journal of Computerized Adaptive Testing

Volume 7 Number 1

February 2019

How Adaptive Is an Adaptive Test: Are All Adaptive Tests Adaptive?

Mark Reckase, Unhee Ju, and Sewon Kim

DOI 10.7333/1902-0701001

The *Journal of Computerized Adaptive Testing* is published by the International Association for Computerized Adaptive Testing
www.iacat.org/jcat

ISSN: 2165-6592

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

Journal of Computerized Adaptive Testing

Volume 7, Number 1, February 2019

DOI 10.7333/1902-0701001

ISSN 2165-6592

How Adaptive Is an Adaptive Test: Are All Adaptive Tests Adaptive?

Mark Reckase, Unhee Ju, and Sewon Kim
Michigan State University

The primary distinguishing feature of computerized adaptive testing (CAT) is that sets of items that are administered to examinees are specifically selected for each examinee during the test process. This implies that examinees who respond differently to items on a test will get different sets of test items. However, if an item bank is small, if there are strict content constraints, or if there is strong exposure control, there might not be much adaptation during the test. Examinees who perform differently might get many of the same test items. The research reported here recommends some statistical indicators of how much adaptation is taking place and shows how these indicators vary for different kinds of adaptive test designs. Guidelines are provided for the values of the statistics indicating that a CAT is strongly adaptive.

Keywords: computerized adaptive test, multistage test, statistical indicators of amount of adaptation

Computerized adaptive testing (CAT) has increased in popularity since its initial large-scale implementation for the Armed Services Vocational Aptitude Battery (ASVAB; see Sands, Waters, & McBride, 1997, for a summary of the development of the CAT-ASVAB). Now CATs are regularly used for licensure/certification tests such as the National Council Licensure Examination (NCLEX; National Council of State Boards of Nursing [NCSBN], 2016) and for state educational assessments (e.g., Minnesota Department of Education, 2017). Even more uses of CAT are likely to be implemented in the future. The term *computerized adaptive test* generally refers to tests that use a computer program to select, administer, and score the individual test tasks presented to a person during the testing session. Item selection is based on the person's previous responses using a trait estimate that is updated before selecting the next test task. Weiss (1974) summarized the strategies that were utilized for adaptive tests up to the date of that report, and many of those strategies continue to be used.

Some adaptive tests have been recreated as new developments, even though they have been used for many years. For example, the Binet intelligence test (Binet & Simon, 1915) was an early application of an adaptive test that used testlets. Currently, there are many variations in CAT design and implementation. The basic form adapts at the individual item level, but these types of CATs are often complicated by exposure control procedures and content balancing. The stratified adaptive tests

(Weiss, 1973), weighted deviation algorithm (Stocking & Swanson, 1993), and shadow test approach (van der Linden & Reese, 1998) are more sophisticated examples of item level adaptation. Some CATs adapt at the level of testlets, such as those that administer a reading passage with a set of associated test items (Wainer, Bradlow, & Du, 2000). Multistage tests (MSTs) adapt using modules of pre-constructed short forms of the test that vary in difficulty (Yan, von Davier, & Lewis, 2016).

In most cases, CATs are designed to adapt the level of difficulty of the test to the trait level of the person taking the test. However, CATs can be designed to adapt to decision-making processes, such as pass/fail on an exam (e.g., Reckase, 1983) or classification into diagnostic categories (Cheng, 2009). CATs also vary on the design of the item bank, whether fixed length or variable length, and on the accuracy of the examinee trait estimation that is desired. These characteristics of CATs affect the way that they function.

Although many design and implementation adaptive testing variations are given the label "CAT," some variations might exhibit more adaptation to the examinees' traits than others. In the worst case, if a CAT has strong exposure control, rigorous content balancing, and a relatively small item bank, there might not be much actual adaptation at all—the severity of the constraints to the test might result in all examinees getting essentially the same test. The possibility that such highly constrained CATs are being used in practice while still being labeled as "adaptive" has led to the present research. The basic questions are: "How adaptive does a test have to be to be labeled as 'adaptive'?" and "How can we tell if a test is meeting the level of adaptation needed to merit the label 'adaptive'?"

The purpose of this paper is to summarize previous work that defines some measures of the amount of adaptation for an adaptive test (Reckase, Ju, & Kim, 2018) and then report the results for using those measures to compare the amount of adaptation of item-level CATs to that of MSTs. Data from an operational MST were also analyzed to show the actual amount of adaptation that was taking place.

The Adaptive Testing Model Used for This Research

The basic assumption of the research reported here was that an examinee has a true location on a latent continuum. The goal was to select a set of test items that would yield the best possible estimate of that location given the practical constraints of item bank size, content balancing, and exposure control. The location on the latent continuum is represented by θ_j for examinee j; and the estimate of the location is represented by $\hat{\theta}_j$. The purpose of the CAT algorithm was assumed to be selecting the best possible set of items for estimating the location of the examinee on the continuum of interest from those items that were available for use.

In the hypothetical ideal case when the location of the examinee on the continuum is known, and there is an infinite item bank with all possible levels of item difficulty available, the best set of items for an n-item test would be that all n items had maximum information at θ_j. For the simple case of the one-parameter logistic model (1PLM), all the items would be selected to have difficulty or location (b) parameters equal to θ_j. The result would be a set of items that had mean b parameters equal to θ_j, with the standard deviation (*SD*) of the bs equal to 0. The mean b parameter for each examinee would also have a correlation of 1.0 with θ_j over examinees. Obviously, this type of CAT never exists in the real world, but the hypothetical case does give some guidance for the types of statistics that can be used to describe the amount of adaptation that occurs in an operational CAT.

In an operational CAT, the analyses described above should show that the mean b parameter for an examinee is related to the final estimated location on the continuum, $\hat{\theta}_j$; that the *SD* of the b parameters for items administered to an examinee is much smaller than the *SD* of the b parameters for the items in the entire item bank; and that the *SD* of the mean b parameter for individual examinees should be about the same as the *SD* of the trait estimates. These three statistical indicators are the basis for the measures of adaptation presented in Reckase et al. (2018).

Proposed Statistical Indicators of the Amount of Adaptation

Reckase et al. (2018) found that a single indicator did not capture all the pertinent information about the amount of adaptation in a CAT. For example, the correlation between the mean *b* parameter for items administered to examinees and the final θ estimate could be high if the item selection algorithm was working properly. However, the mean *b* parameter could be far from the final θ estimate because of limitations in the item bank. That is, the item bank might have a narrow range of difficulty, while the examinees might have a wide range of trait levels. To help interpret the correlation, two other indicators were also developed to account for the spread in difficulty of the items administered to an examinee and the amount of variation of difficulty in the item bank.

The three indicators that were proposed and evaluated in Reckase et al. (2018) were: (1) the correlation between the mean *b* parameter and the final θ estimates for the examinees, $r(\bar{b}_j, \hat{\theta}_j)$; (2) the ratio of the *SD* of the mean *b* parameters for the items administered to the *SD* of the θ estimates for the examinees, $s_{\bar{b}_j}/s_{\hat{\theta}_j}$; and (3) the proportion of reduction of the variance (PRV) of the *b* parameters for the items selected for examinees from the amount of variance of the *b* parameters for all of the items in the bank,

$$\text{PRV} = \frac{s_b^2 - pooled\ s_{b_j}^2}{s_b^2}. \tag{1}$$

These statistics are for CATs using item selection based on minimizing the difference between the *b* parameter for the selected item and $\hat{\theta}_j$. Alternatively, the location of the point of maximum information for item $I(\theta_i^*)$ can be used in place of the *b* parameter computed using the formula provided in Birnbaum (1968), where a_i is the discrimination parameter, b_i is the *b* parameter, c_i is the pseudo-guessing parameter, and D is a scaling constant that makes the logistic function similar to the normal ogive function,

$$\theta_i^* = b_i + \frac{1}{\text{D}a_i}\log\left(\frac{1+\sqrt{1+8c_i}}{2}\right). \tag{2}$$

For the three-parameter logistic model (3PLM), the point of maximum information is slightly higher than the *b* parameter, so the selection of items will be slightly different than selecting on the *b* parameter alone. Reckase et al. (2018) computed the statistics using all of the items administered during a CAT as well as the items from the last half of a CAT and found little difference. In this article, all of the items administered during the CAT were used.

The results of that research showed that the statistical indicators of adaptation were sensitive to item bank size and spread (see Tables 1 and 2 for a summary). The CAT model that was used to develop these tables was maximum information item selection and maximum likelihood θ estimation with a fixed-test length of 30 items. For Table 1, the item bank had normally distributed item difficulties with mean 0 and *SD* = 1. For Table 2, the item difficulties were also normally distributed with mean 0, but the *SD* of the item bank difficulties varied. The results presented in Table 1 show that the statistics improved with increase in item bank size, but all reached high values when the item bank was about 10 times the test length. The results in Table 2 show that the adaptation statistics continued to increase when the *SD* of the *b* parameters for the item bank was larger than the *SD* of the true θs. This is consistent with the need for items more extreme than the true θs to yield accurate estimates of the θs.

Table 1. Evaluation of Adaptation for Different Size Item Banks With a 30-Item Test Length Using Estimated θ

Item Bank Size	Statistic		
	$r(\bar{b}_j, \hat{\theta}_j)$	$s_{\bar{b}_j}/s_{\hat{\theta}_j}$	PRV
50	0.89 (0.01)	0.47 (0.01)	0.62 (0.01)
100	0.92 (0.01)	0.74 (0.02)	0.81 (0.01)
150	0.92 (0.01)	0.79 (0.02)	0.82 (0.01)
200	0.93 (0.01)	0.82 (0.02)	0.82 (0.01)
250	0.93 (0.01)	0.84 (0.02)	0.81 (0.01)
300	0.93 (0.01)	0.85 (0.02)	0.80 (0.01)
350	0.93 (0.01)	0.85 (0.02)	0.79 (0.01)

Note. The values in parentheses are the *SD*s of the statistics over 500 replications.

Table 2. Evaluation of Adaptation for Item Banks With Different *SD*s of *b* Parameters With a 30-Item Test Length Using Estimated θ

Item Bank *SD*	Statistic		
	$r(\bar{b}_j, \hat{\theta}_j)$	$s_{\bar{b}_j}/s_{\hat{\theta}_j}$	PRV
0.1	0.82 (0.01)	0.13 (0.00)	–0.29 (0.07)
0.2	0.83 (0.02)	0.23 (0.01)	0.08 (0.05)
0.3	0.84 (0.01)	0.35 (0.01)	0.21 (0.05)
0.4	0.86 (0.01)	0.47 (0.02)	0.36 (0.03)
0.5	0.88 (0.01)	0.56 (0.02)	0.50 (0.02)
0.6	0.89 (0.01)	0.63 (0.02)	0.55 (0.02)
0.7	0.90 (0.01)	0.70 (0.02)	0.63 (0.02)
0.8	0.91 (0.01)	0.75 (0.02)	0.69 (0.02)
0.9	0.93 (0.01)	0.81 (0.02)	0.74 (0.01)
1.0	0.94 (0.01)	0.85 (0.02)	0.78 (0.01)
1.1	0.94 (0.01)	0.86 (0.02)	0.82 (0.01)
1.2	0.95 (0.01)	0.89 (0.02)	0.84 (0.01)
1.3	0.95 (0.01)	0.91 (0.02)	0.86 (0.01)
1.4	0.95 (0.01)	0.92 (0.02)	0.88 (0.00)
1.5	0.95 (0.01)	0.93 (0.02)	0.89 (0.00)

Note. The values in parentheses are the *SD*s of the statistics over 500 replications.

Reckase et al. (2018) also considered the influence of exposure control on the amount of adaptation. Two exposure-control procedures were investigated: randomesque (randomly selecting an item from the *m* items with the most information) and the Sympson-Hetter procedure (Sympson & Hetter, 1985). The results of simulations of these exposure-control procedures with well-designed item banks showed that exposure control did not have detrimental effects on adaptation. The exception was when the item bank was small and the Sympson-Hetter procedure was used. The PRV statistic was sensitive to the influence of exposure control when the item bank size was not sufficient to support the requirements for the exposure-control procedures (see Table 3).

Table 3. Influence of Exposure Control on Adaptation for Full- and Half-Size Item Banks

Item Bank Size	Exposure Control	Statistic		
		$r(\bar{b}_j, \hat{\theta}_j)$	$s_{\bar{b}_j} / s_{\hat{\theta}_j}$	PRV
Full-size: 306 Items	0	0.95 (0.00)	0.97 (0.01)	0.95 (0.00)
	Randomesque	0.95 (0.00)	0.97 (0.01)	0.95 (0.00)
	Sympson-Hetter	0.96 (0.00)	0.97 (0.01)	0.94 (0.00)
Half-size: 153 items	None	0.96 (0.00)	0.95 (0.01)	0.94 (0.00)
	Randomesque	0.96 (0.00)	0.94 (0.01)	0.93 (0.00)
	Sympson-Hetter	0.94 (0.00)	0.99 (0.02)	0.32 (0.01)

Note. The values in parentheses are the *SD*s of the statistics over 500 replications

The previous research also showed that the level of adaptation for one operational test, the NCLEX (NCSBN, 2016), was very good. The general outcomes of the Kim, Ju, & Reckase (2018) and Reckase et al. (2018) research studies were benchmark values for the statistics that can be used for evaluating the amount of adaptation that occurs in an operational CAT. These benchmark values are shown in Table 4. The values given in the table are for a 30-item CAT based on item selection minimizing the difference between the *b* parameter and $\hat{\theta}_j$.

Table 4. Benchmark Values for Good Levels of Adaptation

Item Response Theory Model	Adaptation Statistic		
	$r(\bar{b}_j, \hat{\theta}_j)$	$s_{\bar{b}_j}/s_{\hat{\theta}_j}$	PRV
1PLM	Low 0.90s	Mid 0.80s	0.80
3PLM	High 0.90s	High 0.70s	0.80

When these indicators of adaptation are obtained for a CAT, examinees with different θ levels will receive tests that match the estimated θ levels very well. Of course, higher values are even better; the one exception is the ratio of the *SD*s. For that statistic, the value 1.0 is optimal, although higher values than 1.0 can be obtained. It is the distance from 1.0 that is important when interpreting that statistic. Values greater than 1.0 can be obtained when the item bank has an unusual distribution of difficulty with many extremely easy and difficult items, but not enough middle-range items. Because that type of item bank is rarely encountered, the benchmark value below 1.0 is given in the table.

Focus of the Current Research

The research reported here focused on the difference in the amount of adaptation that occurs when an MST design is used instead of a test design that adapts at the level of the item or testlet. MST designs are typically made up of a fixed-form routing test that is used to select the second-stage tests (or modules) of varying difficulty. There might be additional routing decisions to select third-stage tests (or modules) to match the capabilities of the examinees.

MST designs are attractive because they allow formal review of the sets of items that make up the modules comprising the stages of the test to determine if they meet content specifications, or if

there are item selection flaws such as cluing or nearly identical content (e.g., Yan et al., 2016). However, MSTs reduce the amount of adaptation that occurs because examinees take the same routing test and there are only a limited number of second- and later-stage modules. For example, if the test has what is called a 1-3 design, one module is at the first stage (the routing test) and three modules are at the second stage. In this case, there are only three possible levels of adaptation. All examinees routed to the same second-stage module will be adapted to their trait level in the same way.

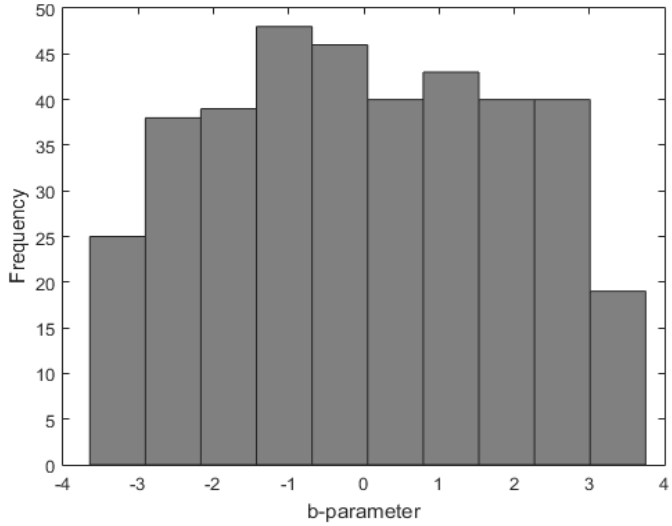
Simulation Studies

The intent of this research was to have a fair comparison between the level of adaptation of MSTs and item-level CATs. To accomplish this goal, a test length of 40 items was selected for both test designs because it is typical of the length of tests used in educational settings, although that is not the case for certification/licensure where the tests are typically much longer. Both the item-level adaptive test and the three-stage tests were designed using a 40-item fixed test length. A standard normal θ distribution was assumed for the examinee population for all simulations.

Given the test length and the distribution of examinees, procedures for the item bank design presented in Reckase (2010) and based on the 1PLM were used to design the desired item bank for the item-level CAT. The item bank had 378 items that were distributed as shown in Figure 1. A total of 1,000 *b* parameters for items were generated from a standard normal distribution to create a master bank. The items were selected from the master bank to match the requirements for the 378-item bank. This was done so that the distributions of items within bins would approximate the distributions that might be obtained in a real situation (see Table 5).

The item-level CAT that was used as a basis for comparison in the study had an initial trait estimate of 0.0, item selection using maximum information at the current estimate of θ, and maximum likelihood θ estimation. At the beginning of the test, the θ estimate was increased by 0.7 after a correct response and reduced by 0.7 after an incorrect response, until both correct and incorrect responses were present in the response pattern (Reckase, 1975). After both correct and incorrect responses were present, the maximum likelihood estimate was obtained and used for the selection of the next item.

Figure 1. Histogram of an Optimal Item Bank for a 40-Item CAT



There does not seem to be an optimal design for MSTs in the research literature, so two design approaches for the 1-3-3 MST were used for this research. The first design approach assumed that the goal was to have equal numbers of examinees administered the modules at each stage and that classification/routing accuracy was important for the early stages of the test. For this design, at the first and

second stages the items were selected to have maximum information at the routing decision points. Because the first stage was used to classify into three different second-stage modules, 20 items were used for this stage. The second- and third-stage modules had 10 items in each module.

Table 5. Item Bank Design for the CAT

Range of *b* for Bins	Frequency	Mean *b* Parameter
$-3.9 \geq b > -3.3$	7	−3.46
$-3.3 \geq b > -2.7$	28	−2.96
$-2.7 \geq b > -2.1$	31	−2.42
$-2.1 \geq b > -1.5$	34	−1.81
$-1.5 \geq b > -0.9$	35	−1.15
$-0.9 \geq b > -0.3$	36	−0.61
$-0.3 \geq b > 0.3$	36	−0.02
$0.3 \geq b > 0.9$	36	0.60
$0.9 \geq b > 1.5$	35	1.20
$1.5 \geq b > 2.1$	34	1.75
$2.1 \geq b > 2.7$	31	2.37
$2.7 \geq b > 3.3$	28	2.97
$3.3 \geq b > 3.9$	7	3.46
Total	378	−0.01

The decision points were selected so that equal numbers of examinees would be routed to each module. The items for Stage 3 were selected to give approximately equal information over the range of θs routed to the modules in that stage, taking into account the amount of information obtained from the previous two stages. The summary statistics for the distributions of items in each model for this design approach are given in Table 6. All items were selected from the operational bank used for the item-level CAT.

Table 6. Summary of *b* Parameter Distributions by Stage for the Equal Distribution of Examinees Design

Stage	Difficulty Level	Number of Items	Routing Points	Mean	*SD*	Min	Max
First stage		20	−0.44, 0.44	−0.00	0.43	−0.49	0.51
Second stage	Easy	10	−0.69	−0.70	0.02	−0.73	−0.66
	Medium	10	−0.28, 0.26	−0.01	0.29	−0.31	0.31
	Difficult	10	0.68	0.67	0.04	0.60	0.75
Third stage	Easy	10		−3.20	0.10	−3.34	−3.08
	Medium	10		−0.96	1.02	−1.54	1.08
	Difficult	10		3.40	0.16	3.18	3.75

The second design approach assumed that the goal was to obtain the same amount of information for all examinees (i.e., equal precision) and that the number of examinees taking each module was not of concern. To achieve this goal for a 1-3-3 design, each module at the second and third stage was designed to have uniform information over a fixed range of θ.

The easy modules in Stages 2 and 3 were designed for the θ range from –3 to –1; the medium modules in Stages 2 and 3 were designed for the range from –1 to 1; and the difficult modules in Stages 2 and 3 were designed for the range from 1 to 3. In this case, the routing test was designed to give uniform information from –3 to 3. The decision points used for routing were –1 and 1 in all cases because these were the boundaries between the ranges used to design the modules. The summary statistics of the distributions of items in each module for this design are given in Table 7.

Table 7. Summary of *b* Parameter Distributions by Stage for the Uniform-Information Design

Stage	Difficulty Level	Number of Items	Routing Points	Mean	*SD*	Min	Max
First stage		20	−1, 1	0.12	2.68	−3.29	3.75
Second stage	Easy	10	−1	−1.47	1.78	−3.34	0.59
	Medium	10	−1, 1	0.07	1.78	−2.21	1.84
	Difficult	10	1	1.54	1.76	−0.49	3.34
Third stage	Easy	10		−1.41	1.38	−3.37	−0.04
	Medium	10		−0.14	1.56	−1.94	1.59
	Difficult	10		1.40	1.70	−0.31	3.14

The distributions of *b* parameters were quite different for the two designs. To achieve uniform information, the distribution of *b* parameters needed to be almost U-shaped so that the information did not accumulate through overlap in the middle of the target range. As a result, the *SD*s of the *b* parameter distributions were large for the uniform-information design. In contrast, the classification accuracy design had low *SD*s for the *b* parameters because information was accumulated at the decision points to maximize classification accuracy.

To compare the amount of adaptation that was present when each of the testing designs was used, test administrations were simulated assuming that the true θ level had a standard normal distribution. The measures of adaptation were computed from simulated samples of 500 examinees from the distribution. Then, the process was replicated 100 times so that the amount of variation in the statistical indicators could be determined. Additionally, several other statistics were computed to check whether the simulations were performing as expected.

Real Data Analysis

The real data analysis of level of adaptation was performed using data from the Early Childhood Longitudinal Study, Kindergarten Class of 1998 to 1999 (ECLS; National Center for Education Statistics [NCES], 2005) that assesses the mathematics, reading, and science performance of third-grade students. All the tests used a two-stage test with a 1-3 design. For mathematics, the first-stage test had 17 items, and each second-stage test had 23 items. For reading, the first-stage test had 15 items. The second-stage tests had numbers of items varying from 24 to 38. For science, the first-stage test had 15 items, and each second-stage test had 20 items.

Results

Simulation Studies

Before presenting the evaluations of the amount of adaptation for the item-level CAT and the two MST designs, some information is provided to evaluate the functioning of those tests. This in-

formation includes the means and *SD*s of the true and estimated θs, the RMSE for the θ estimates, and the correlation between the true and estimated θs. These results are presented in Table 8.

Table 8. Summary Statistics for θ Estimates From the Different Test Designs

	True θ		Estimated θ			
Test Design	Mean	*SD*	Mean	*SD*	RMSE	$r\left(\theta,\hat{\theta}\right)$
Item-level CAT	0.01	1.02	0.01	1.07	0.33	0.95
Classification 3-stage MST	0.01	1.02	0.02	1.09	0.37	0.94
Uniform information 3-stage MST	0.01	1.02	0.01	1.09	0.39	0.93
Fixed-form test	0.01	1.02	0.02	1.12	0.42	0.93

Note. RMSE = root mean squared error.

The descriptive statistics were similar for the three test designs, but the correlation between the true and estimated θs was slightly higher for the item-level CAT. In all cases, the estimated θs had slightly larger *SD*s than the true θs due to the addition of estimation error. These results are not surprising because a non-adaptive 40-item linear test would yield a similar pattern of results. The last row of the table shows the descriptive statistics for a fixed 40-item test developed by randomly sampling the items from the master item bank. Note that it had a substantially larger RMSE than the item-level CAT.

The simulation results were also analyzed to determine whether the routing rules for the three-stage tests worked as expected, and whether the information plots for the uniform-information approach to the three-stage test gave the expected results. The proportions of examinees routed to each module at each stage are shown in Table 9. The routing, designed to give equal proportions in each second- and third-stage module, gave the expected results. They were not exactly equal to 33% for each module because of the discrete nature of the θ estimates that are obtained from the 1PL model—there are only as many unique trait estimates as there are number-correct scores. In this case, the number-correct scores for the routing test ranged from 0 to 20. When frequencies are allocated to only 21 θ estimates, it is usually not possible to divide the distribution into exactly equal thirds.

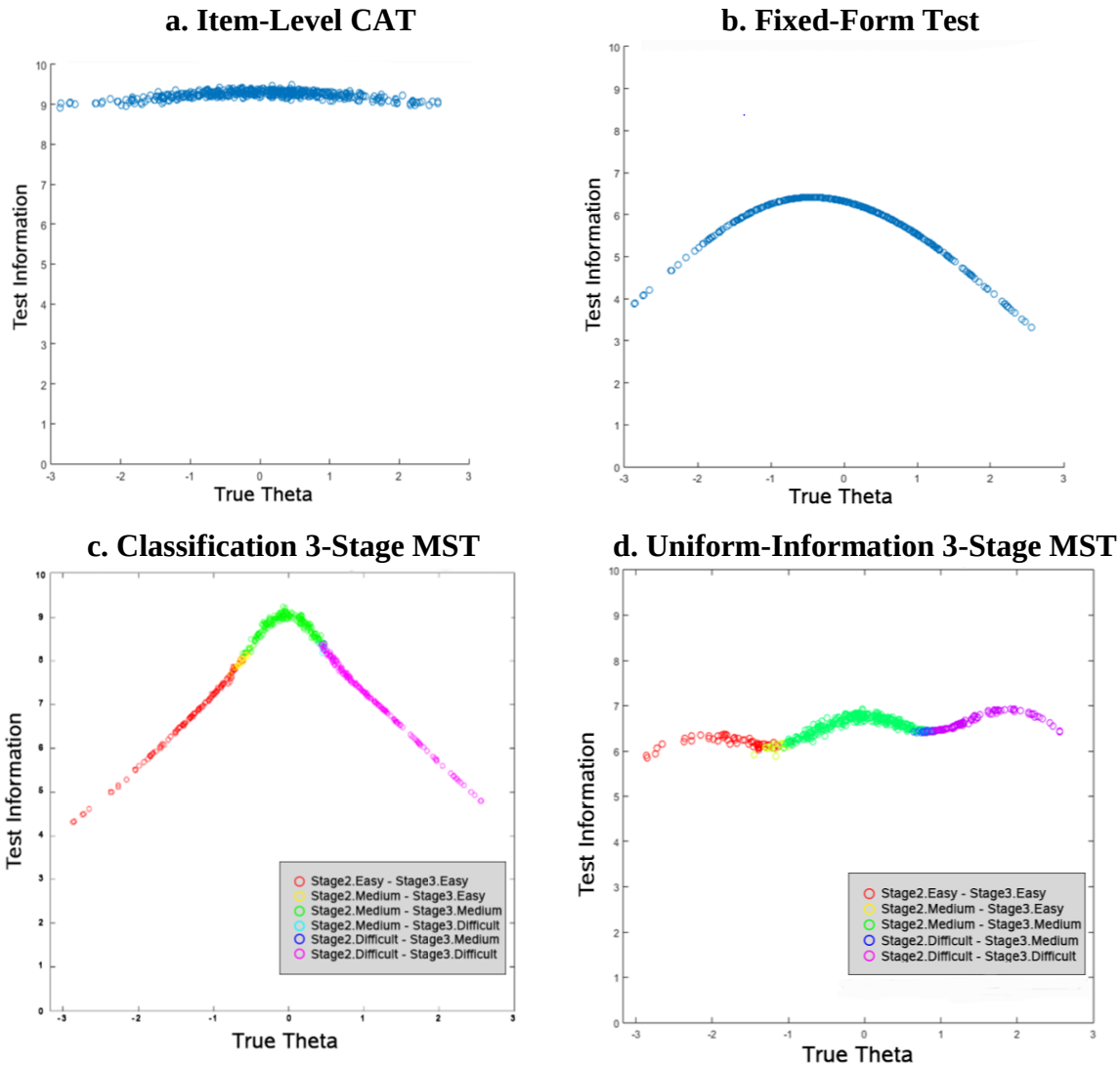
Table 9. Proportion of Examinees Routed to Each Module for the Equal-Proportion Routing and the Uniform-Information Routing

	Module		
Stage	Easy	Medium	Difficult
Equal-proportion routing			
Stage 2	29.5	33.3	37.3
Stage 3	29.8	35.4	34.7
Uniform-information routing			
Stage 2	16.0	54.8	29.1
Stage 3	17.4	61.3	21.3

The uniform-information routing resulted in more examinees routed to the medium-level modules than the easy or difficult modules. This was a result of using equal ranges on the θ scale for this routing. The density of the normal distribution varied substantially over those equal-length ranges.

Plots of the average information at points along the θ scale for the different test designs are shown in Figure 2. The plot shows the information for the item-level CAT, the fixed-form test sampled from the item bank, the equal-proportion routing MST, and the uniform-information MST. The graphs show that the uniform-information MST did give approximately uniform information over the θ range. The equal-proportion MST had a highly peaked information graph because of the accumulation of information in the middle range resulting from the design of the routing test and second-stage modules. The modules were designed to minimize misclassifications into the next higher-level modules, so the information was concentrated at the cut scores. The uniform-information MST had the information spread over the full range of each interval.

Figure 2. Average Test Information Conditional on θ for the Test Designs Used in the Study



These initial analyses showed that the designs of the MSTs worked as desired. The uniform-information MST gave approximately uniform information. The equal-proportion MST assigned approximately equal proportions of examinees to each module. The item-level CAT and the fixed-form test are provided for comparison purposes.

Table 10 shows the results of the adaptive analysis for the different test designs. The item-level CAT had high levels of adaption. All statistics were above the benchmark values, which was to be expected because a well-designed item bank was used, and the test was of sufficient length to give good adaptation. The MSTs had adaptation statistics that were below the benchmark values. The reason for the low values was due to substantial variation in difficulty within modules for the MSTs. The uniform-information MST was also somewhat less adaptive than the equal-proportion MST. The largest difference was for the PRV statistic, resulting from the amount of within-module variation in difficulty for the different designs. This variation was greater for the uniform-information MST due to the requirements of the design, which had items spread over the θ range covered by the module. The range of difficulty was necessary to achieve uniform information, while the test design that gave equal proportions to each module had items concentrated at decision points.

Table 10. Adaptation Statistics for the Different Test Designs

Test Design	Statistic		
	$r(\bar{b}_j, \hat{\theta}_j)$	$s_{\bar{b}_j}/s_{\hat{\theta}_j}$	PRV
Item-level CAT	0.96 (0.00)	0.98 (0.01)	0.94 (0.00)
Classification 3-stage MST	0.88 (0.01)	0.62 (0.01)	0.50 (0.01)
Uniform information 3-stage MST	0.84 (0.01)	0.62 (0.01)	0.33 (0.00)
Fixed-form test	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)

Note. Empirical *SD*s (i.e., standard errors) are in parentheses.

Real Data Analysis

The results from the real data analysis of the ECLS (NCES, 2005) are shown in Table 11. This table contains results for the two-stage tests for mathematics, reading, and science. Because the ECLS uses the 3PLM model for θ estimation, the adaptation statistics were computed based on both the *b* parameters for the items and the point of maximum information on the θ scale for each item. For the ECLS, the results are almost identical for the two indicators of item difficulty. This might be because many items had low or zero *c* parameters, so there was not much shift in the location of maximum information.

Of the three ECLS content areas, mathematics showed the greatest amount of adaptation, and reading showed the least. In all cases, however, the amount of adaptation was below the benchmark values identified for the item-level adaptive tests shown in Table 4. The adaptive statistics for the mathematics test were similar to those for the simulated MST when equal-proportion routing was used.

The reading and science tests do not show as high a level of adaptation. For the reading and science tests, the results are probably a result of the routing rules. For those tests, at least 50% of the examinees were routed to the middle-level test in the second stage. Thus, there was not much adaptation. Also, for the reading test, the difference in mean difficulty for the modules in the second stage was fairly small—about .30 on the θ scale. That difference was less than the *SD* of the *b* parameters within a module. In addition, the reading test selects a passage and all the items that go with it. The passages typically have a range of difficulty of items that are connected to them, resulting in a low PRV statistic.

Table 11. Adaptation Statistics for the ECLS Test

Test and Item Statistic	Adaptation Statistic		
	$r(\bar{\zeta}_j, \hat{\theta}_j)$	$s_{\bar{\zeta}_j}/s_{\hat{\theta}_j}$	PRV
Mathematics			
b parameter	0.86	0.74	0.47
Maximum information point	0.86	0.73	0.47
Reading			
b parameter	0.76	0.44	0.25
Maximum information point	0.76	0.46	0.26
Science			
b parameter	0.83	0.50	0.42
Maximum information point	0.83	0.49	0.40

Note. The symbol "ζ" represents either the point of maximum information or the *b* parameter

Discussion and Conclusions

The focus of this research was to quantify the difference in amount of adaptation provided by an MST as compared to an item-level CAT. Three statistical descriptors were suggested as ways for quantifying the amount of adaptation: (1) the correlation between the estimated θs and the mean item location indicator[1] for the items administered to each person; (2) the ratio of the *SD* of the mean item location indicators to the *SD* of the θ estimates; and (3) the PRV of the location indicators administered to an individual, compared to the variance of the item location indicators for the full item bank. These statistical descriptors provide slightly different information about the amount of adaptation. Together, they provide objective information about the matching of the item selection to the level of performance for everyone taking the test.

Some initial results from Reckase et al. (2018) were provided to serve as a basis for comparison with the present study. The results were from a simulation study of the function of an item-level CAT selecting items to minimize the difference between the *b* parameter and $\hat{\theta}_j$. Those results provided benchmark values for the amount of adaptation that can be achieved from an item-level CAT; results from the NCLEX examination showed that the level of adaptation specified by the benchmarks can be achieved in practice (Reckase et al., 2018).

The research reported here was a comparison between the amount of adaptation provided by an item-level CAT and that provided by MSTs developed from the same item bank. Two MST designs were used—one designed to have equal numbers of administrations using each module in the design and one designed to have uniform information for all examinees. The results for the adaptation statistics indicated that the MST provided less adaptation than an item-level CAT and that the MSTs with equal proportions routed to each module had slightly better adaptation than the uniform-information MST.

[1]Point of maximum information or *b* parameter.

Real data analyses using data from the ECLS two-stage tests (NCES, 2005) showed variation in the amount of adaptation by content area. The mathematics test showed the most adaptation; the reading test the least, with the science test between the other two. The mathematics test showed results that were similar to the simulation results. It appears that the difference in the amount of adaptation for the different tests was a result of the routing rules used and the designs of the modules. Routing rules that resulted in equal proportions being routed to the modules, and modules that were distinctly different in difficulty would increase the amount of adaptation.

Part of the motivation for this research was to help users of tests determine whether a test called "adaptive" really is adaptive. The results show that there was variation in the amounts of adaptation observed. Those differences in amount of adaptation resulted in differences in the error in estimation of the trait as shown by the RMSE. These differences are due to differences in the amount of information provided by the adaptation algorithms.

References

Binet, A. & Simon, T. (1915). *A method of measuring the development of intelligence of young children* (3rd edition) (C. H. Town, Trans.). Chicago, IL: Chicago Medical Book Co.

Birnbaum, A. (1968). 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*, 397-479. Reading, MA: Addison-Wesley.

Cheng, T. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika, 74*(4), 619-632. *CrossRef*

Kim, S., Ju, U., & Reckase, M. D. (2018, April). *Evaluating indicators of amount of adaption to 3PL computerized adaptive test.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Minnesota Department of Education (2017). *Technical manual for Minnesota standards-based accountability and English language proficiency assessments: For the academic year 2015–2016.* Retrieved from file:///C:/Users/early/Downloads/20152016%20Technical%20Manual%20for%20Minnesota's% 20MCA%20and%20MTAS%20Assessments.pdf

National Center for Education Statistics (2005). *Early childhood longitudinal study, kindergarten class of 1998 to 1999 (ECLS-K): Psychometric report for the third grad*e. (NCES 2005-062). Washington, DC: US Department of Education, Institute of Education Sciences. Retrieved from https://nces.ed.gov/pubs2005/2005062.pdf

National Council of State Boards of Nursing (2016). *NCLEX-RN examination: Test plan for the National Council Licensure Examination for Registered Nurses*. Retrieved from https://www.ncsbn.org/RN_Test_Plan_2016_Final.pdf

Reckase, M. D. (1975, April). *The effect of item choice on ability estimation when using a simple logistic tailored testing model*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 237–257). New York, NY: Academic Press

Reckase, M. D. (2010). Designing item pools to optimize the functioning of a computerized adaptive test. *Psychological Test and Assessment Modeling, 52*(2), 127-141. *CrossRef*

Reckase, M. D., Ju, U., & Kim, S. (2018). Some measures of the amount of adaptation for computerized adaptive tests. In M. Wiberg, S. Culpepper, R. Janssen, J. Gonzalez, & D. Molenaar (Eds.), *Quantitative psychology: The 82nd annual meeting of the Psychometric Society*. Cham, Switzerland: Springer International Publishing.

Sands, W. A., Waters, B. K., & McBride, J. R. (Eds.) (1997). *Computerized adaptive testing: From inquiry to operation.* Washington, DC: American Psychological Association.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17*(3), 277-292. *CrossRef*

Sympson, J. B., & Hetter, R. D. (1985, October). *Controlling item exposure rates in computerized adaptive testing*. Paper presented at the annual meeting of the Military Testing Association. Navy Personnel Research and Development Center, San Diego, CA.

van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, *22*(3), 259–270. *CrossRef*

Wainer, H. Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.) *Computerized adaptive testing theory and practice* (pp. 245-269). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Weiss, D. J. (1973). *The stratified adaptive computerized ability test*. (RR 73-3). Minneapolis, MN: Psychometric Methods Program, Department of Psychology, University of Minnesota. Available at http:// http://iacat.org/biblio

Weiss, D. J. (1974). *Strategies of adaptive ability measurement* (RR 74-5). Minneapolis, MN: Psychometric Methods Program, Department of Psychology, University of Minnesota. Available at http:// http://iacat.org/biblio

Yan, D., von Davier, A. A., & Lewis, C. (Eds.) (2016). *Computerized multistage testing: Theory and applications.* Boca Raton, FL: CRC Press.

Authors Address

Mark Reckase, Michigan State University, 450 Erickson Hall, 620 Farm Lane, East Lansing, MI 48824 U.S.A., reckase@msu.edu; Sewon Kim, kimsewon@msu.edu; Unhee Ju, juunhee@msu.edu