ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(4); 358-373

Mixed Adaptive Multistage Testing: A New Approach

Anthony RABORN * Halil Ibrahim SARI **

Abstract

Computerized adaptive testing (CAT) and computerized multistage testing (CMT) are two popular versions of adaptive testing with their own strengths and weaknesses. This study proposes and investigates a combination of the two procedures designed to capture these strengths while minimizing the weaknesses by replacing the standard MST routing module with a CAT-based, item-level routing module. A total of 3000 examinees were simulated from a truncated normal distribution with bounds at -3 and 3, and a simulation study was conducted. Simulation results indicate that the new method provides some efficiency improvements over traditional MST when both routing modules are the same size, and when the item-level routing module is larger, the improvements are greater. The study showed that the proposed test administration model could be used to measure student ability, meaning that our new method resulted in lower mean bias, lower RMSE, and higher correlation than traditional MST. An R package built from the code used for this paper is also introduced in the supplementary file. The limitations of the study and recommendations for future research are also presented.

Key Words: Computerized adaptive test, multistage adaptive test, simulation, R, mixed adaptive test.

INTRODUCTION

There are two popular adaptive testing approaches: computerized adaptive testing (CAT) (Weiss & Kingsbury, 1984) and multistage testing (MST) (Luecht & Nungester, 2000). CAT is more widely known and more often used; in this approach, an examinee receives an item typically at medium difficulty level (e.g., maximizing information at the theta level of 0) and, based on his/her response to previous item(s), the item selection algorithm selects the next item from a large item pool. This continues until the examinee completes the test. A well-known advantage of CAT is allowing all test takers to work own personalized test producing high measurement accuracy in ability estimation (Yan, von Davier, & Lewis, 2016). In MST, however, the test has a panel design describing how different sets of items (e.g., 10 items) called modules are grouped into different stages. In stage one, there is typically one module called the routing module. In subsequent stages, there are several modules at different difficulty levels (e.g., easy, medium, and hard difficulty modules). An MST can be comprised of several stages and a different number of modules in each stage. For example, a 1-3-4 design has one module in stage one, three modules in stage two, and four modules in stage three. The working principle of MST is as follows. An examinee initially receives a set (e.g., 5 or 10 items) typically at the medium difficulty level. Based on the examinee's performance on this routing module, the module selection algorithm selects the next module from the next stage (Luecht, Brumfield, & Breithaupt, 2006). This continues until the examinee completes all the stages. The main difference between these two types of test administrations is that there is item-level adaptation in CAT but module-level adaptation in MST. Each has its own advantages and disadvantages.

MST has the disadvantage of being somewhat less efficient than CAT, meaning that CAT results in better theta estimates with lower standard errors than MST in many circumstances (Luecht & Sireci, 2012). This is due to item level adaptation feature of CAT. However, many common item-level adaptation schemes use maximum item information as the criterion for item selection, meaning the first few items selected by maximum information have higher exposure rates than later items; this can

* Supervisor, Accountability, Research, and Measurement, Pasco County Schools, Florida-United States of America, araborn@pasco.k12.fl.us, ORCID ID: 0000-0002-8083-4739
** Assoc. Prof., Kilis 7 Aralik University, Muallim Rifat Faculty of Education, Kilis-Turkey, hisari87@gmail.com, ORCID ID: 0000-0001-7506-9000

To cite this article:
Raborn, A., & Sarı, H. (2021). Mixed adaptive multistage testing: A new approach. *Journal of Measurement and Evaluation in Education and Psychology, 12*(4), 358-373. doi: 10.21031/epod.871014

Received: 30.01.2021
Accepted: 14.10.2021

be alleviated by modifying the item-level adaptation to choose from more than just the most informative item (Barrada, Olea, Ponsoda, & Abad, 2008). Another advantage of CAT over MST is that CAT allows for both varying and fixed test lengths, but the traditional MST is a fixed test length exam.

MST has the advantage of permitting test takers to answer or change answers to any question within the current module at any time while allowing for tests constructed to meet specific content and length requirements. This is an advantage because allowing response change provides having lower standard errors for the ability estimates, especially for students with higher abilities (Liu, Bridgeman, Gu, Xu, & Kong, 2015). Another advantage of MST is that it allows higher levels of control to test developers. This means that the developer of the test can place items into a module and easily keep track of content balancing, item usage, test length, or other statistical and non-statistical test requirements. However, these issues can sometimes be a problem in CAT, especially when there is a limited number of items from a content area in the item pool (Robin, Steffen, & Liang, 2016). Due to test assembly occurring prior to the test administration in MST, there is always greater expert control over item order and content area in this format (Sari & Huggins-Manley, 2017).

The better efficiency from CAT comes at the cost of the complex algorithms needed for the item-level adaptation, which MST avoids by having fewer adaptation points. This is because there are *n*-1 adaptation points in CAT, where *n* is the total test length as opposed to *k*-1 adaptation points in MST, where *k* is the number of stages. Having fewer adaptation points in MST has its own cost. For example, recovery of ability estimates becomes a difficulty when examinees are misrouted (e.g., incorrectly routed) through the modules. Previous research has shown that the initial routing stage has a major influence on the accuracy of final theta estimates, particularly in two-stage tests (Kim & Plake, 1993). Since the routing module provides the provisional theta estimate for the next modules, the routing module should include items from a wider range of difficulties. This means that it should maximize module level test information function at a wider theta range for test takers having different theta levels. Otherwise, it would be difficult to make better initial estimates for all test takers. A poorly designed routing module (e.g., with a very low maximum value for the test information function and/or very difficult or easy items) can place an examinee in the incorrect module in the subsequent stage. This would result in dramatic changes in the pathways one draws during the test. Consequently, it might be difficult or impossible to obtain less bias for the final theta estimate (Sari, Yahsi-Sari, & Huggins-Manley, 2016). As the number of stages increases, this influence is reduced, but practical considerations limit the number of stages that can be created and administered. Furthermore, previous studies showed that the reduction in estimation error provided by increasing the number of stages is modest (Patsula, 1999; Zenisky, Hambleton, & Leucht, 2010). A solution to establish better measurement accuracy after the routing module would be to increase the number of items in the routing module, but this would lead to an increase in the number of retired items after the test. This is because routing items are seen by all examinees and therefore reach maximum exposure rate.

Prior Attempts to Combine CAT and MST

A review of the literature showed that there was one other study that compared a proposed combination of CAT and MST. Wang, Lin, Chang, and Douglas (2016) performed three simulation studies investigating Hybrid Computerized Adaptive Testing, which used MST for the initial items and CAT for the subsequent items and compared it to traditional MST. Their hybrid test starts with MST (e.g., module-level adaptation) for the first two adaption points then uses CAT (e.g., item-level adaptation) for the remaining adaptation points. The first two simulations varied the proportion of items in the test that fell under the MST framework from 1/3rd of the test length to 5/6th of the test length and investigated six common MST designs, while the last simulation compared the two best designs from the first two simulations to two CAT and two MST designs. Their results indicated that, with two and three stages of various lengths, stage designs, and proportion of items in the MST stages, the hybrid designs (i.e., the combination of MST and CAT) perform as well or better than the traditional CAT design in terms of bias and RMSE and better than the studied MST designs in terms of RMSE.

In their study, the authors approached the problem primarily from the perspective of CAT (e.g., starting with MST and switching to CAT). In addition, their first simulation only used the two-stage 1-4 panel design for the MST comparison, and none of the three simulations fully compared the efficacy of the hybrid design to traditional MST designs. Thus, no single simulation included all of the factors manipulated in the study. This study, on the other hand, aims to follow the MST framework. Also, our study uses only three-stage designs but investigates the effect of MST design complexity and overall test length in the proposed hybrid design. The different emphases, as well as the different strengths of the approaches, lend credence to the investigation of ma-MST as an alternative to traditional MST and the other hybrid designs.

Purpose of the Study

In order to increase initial measurement accuracy while maintaining item exposure limits and allowing examinees to change answers within certain modules, we propose combining the CAT and MST methods into a single test administration. We called this new administration type as a mixed adaptive multistage test (ma-MST). The ma-MST will start with a CAT-based routing module (e.g., item-level adaptation) and obtain a provisional theta estimate. Then, this provisional theta estimate will be used to select the next MST-based stage. This means that the exam will start with CAT and switch to MST. We aim to bring MST closer to the efficiency of CAT while maintaining the aforementioned benefits of MST. By combining the methods in this way, the likelihood of misrouting can be reduced by the more accurate measure of ability after administrating items with item-level adaptation. As a result, this would result in a lower bias for the estimations of ability by the end of the test, while still allowing for easier control of item exposure rates, content balancing compared to the traditional MST, and allowing examinees the ability to change their answers in the later stages.

A New Approach: Mixed Adaptive Multistage Test (ma-MST)

Using R (R Core Team, 2016) and the R package “caMST” (Raborn, 2018), we investigated the efficacy of using item-level adaptation to route individuals to further modules. This new test format, mixed adaptive multistage test (ma-MST), is similar to a traditional MST in that it has a specific number of stages administered. However, the number of potentially administered tests is greater than in MST but less than in CAT because individuals would share panels of items depending on their ability estimates after seeing potentially different items in the CAT-based routing module.

This new method has much of the same test assembly processes as typical multistage tests and utilizes automated test assembly (ATA) to create each panel at each stage. In theory, ma-MST has similar item pool requirements as both CAT and MST. Item exposure concerns also remain and should be handled as appropriate for the use of the test (e.g., Reckase, 2010; van der Linden, 2000). In order to simplify the initial investigation of this method, there will not be any exploration of overall item exposure differences between CAT, MST, and ma-MST. This means that item exposure concerns will be ignored in favor of focusing on determining the accuracy of the different methods in their ability estimates.

In this study, the hybrid approach (e.g., ma-MST) will include a larger proportion of items selected with item-level adaptation points than in modules (e.g., resembling CAT). The ma-MST will also include a larger proportion of items in a module than selected with item-level adaptation points (e.g., resembling traditional MST). The primary goal of this study is to investigate the efficiency of the ma-MST, and what happens to the estimated theta parameters when the hybrid model resembles CAT and traditional MST. The expectation is that ma-MST would have lower bias and RMSE, higher correlation in the final theta estimations, especially as the CAT proportion increases.

For this study, we had two main research questions to answer:

1. How is the test efficiency (Bias, RMSE, and Correlation) be impacted when;
   a. CAT proportion (1/6, 1/3, and 2/3),

   b. CAT item selection method (MFI, random selection),
   c. MST designs (1-2-2, 1-2-3, and 1-3-3),
   d. Test length (18 and 30 items) are varied in mixed approach simulations?

2. How will the test efficiency (Bias, RMSE, and Correlation) be impacted on the mixed adaptive and traditional MST under the combination of the levels of test length and MST design?

METHOD

We performed a simulation study to test the efficacy of the ma-MST against a traditional MST using the “caMST” package in R. The annotated R codes that demonstrate how to use the package to replicate the methods described here are provided in the supplementary file. We held constant the following factors: a) the number of stages (held at 3), b) the number of panels (3 parallel panels), c) the module selection or routing procedure (select the module with the maximum Fisher information [MFI] at the provisional theta), d) the initial ability estimate (held at $\theta_{\text{initial}} = 0$) and e) the provisional and final ability estimation procedures (expected a posterior [EAP], as commonly used in previous studies (Briethaupt & Hare, 2007; Luecht et al., 2006). The factors that we varied were the MST panel design, total test length, the fraction of the CAT routing module to the total test length, and the item selection procedure for CAT for a total of thirty-six conditions (see Table 1 for the levels). In addition to the ma-MST factors above, we used a traditional MST procedure as a baseline for each module design and test length.

Table 1. Simulation Study Conditions and Levels

Factor	Number of Levels	Levels
Panel Design	3	1-2-2 1-2-3 1-3-3
Test Length	2	18 items 30 items
CAT Module Length (fraction of overall test length)	3	1/6 1/3 2/3
Routing Module Item Selection	2	-Maximum Fisher information (Random 1 MFI) -Random selection from 5 items with Maximum Fisher information (Random 5 MFI)
	3x2x3x2=36	

The item parameters were based on a real Armed Services Vocational Aptitude Battery (ASVAB) military test used in Armstrong, Jones, Li, and Wu (1996). The simulated item bank had 450 multiple-choice items from four different content areas. In this study, in the 30-item condition, there were 10, 11, 4, and 5 items in content areas 1 through 4, respectively. For the 18-item condition, they are set to 6, 6, 3, and 3 items, respectively. The item parameters and the number of items for each content area in the original study were given in Table 2.

Table 2. Item Characteristics per Content Area

Content Area	*a*		*b*		*c*	
(Number of items)	Mean	SD	Mean	SD	Mean	SD
Content 1 (*n* = 150)	1.079	.409	-.467	1.179	.210	.095
Content 2 (*n* = 165)	1.128	.438	-.154	1.033	.200	.104
Content 3 (*n* = 60)	1.092	.538	-.025	.815	.203	.084
Content 4 (*n* = 75)	1.237	.383	-.014	.678	.162	.080

Armstrong et al. (1996)

A total of 3000 examinees were simulated from a truncated normal distribution with bounds at -3 and 3. Response patterns were generated according to Birnbaum's (1968) three-parameter (3PL) model in R. We used the EAP estimator (Bock & Mislevy, 1982) from the "mstR" package (Magis, Yan, & von Davier, 2017) with the prior distribution *N*(0, 1) for all ability estimation. The IBM CPLEX program (ILOG, 2006) was used to construct the various modules in stages 2 and 3, and three essentially (although not strictly) parallel panels (i.e., the same number of items from the different content areas and similar in difficulty level). The items that were not used in these stages were treated as a mini item bank for the CAT and, depending on the test length and CAT proportion, the computer algorithm selected items from this bank consisting of the items remaining after the ATA. The bottom-up strategy was used when building the panels. The content distributions in the modules across the different test length and panel design conditions were given in Supplementary Tables 1, 2, and 3, under the 1/6, 1/3, and 2/3 CAT conditions, respectively. The panel-level test information across the CAT proportion and test length conditions were given in Supplementary Figure 1. For the modules in stages two and three, the module level information function was maximized at the fixed theta points of $\theta = -1$, $\theta = 0$, and $\theta = 1$ for the easy, medium, and hard modules in the conditions, respectively. In the baseline condition (e.g., traditional MST), the routing module was maximized at the theta point of 0.

Again, for the conditions with the CAT-based routing module, the items were selected from the pool of items that were not used for the modules. Then, for the random 1 MFI condition, the most informative item which fit the content area specification mentioned above was selected. For the random 5 MFI condition, a random item from the five most informative items which fit the content area specification was selected. This process was repeated after each item, updating the information function with every answer choice, until the simulated respondent answered the maximum number of items for the routing module.

The working principle of ma-MST simulation was as follows. In each design (e.g., 1-2-2, 1-2-3, or 1-3-3), if the CAT proportion was 1/6, and the total test length was 18, the computer tailored three items (1/6 of the 18 items) to the individual based on their responses in the first stage (e.g., item-level adaptation), and tailored 15 items in the two remaining stages (e.g., module-level adaptation). If the total test length was 30 and CAT proportion was 2/3, simulated individuals were administered 20 CAT-based items in the first stage and 10 total MST-based items in the second and third stages. This indicates that under the same total test length, as the CAT proportion increases, more items are administered at the item level.

To determine the efficiency of the tests within these conditions, we calculated mean bias, root mean squared error (RMSE), and Pearson correlations between true theta and estimated theta. It is important to note that each overall statistic was calculated for across the 3000 examinees for a replication (e.g., iteration) and averaged across 100 replications.

For the results, we ran a four-way factorial ANOVA separately for each of the outcomes, keeping the highest-order interaction terms in each case. To determine the magnitude of any experimental effects, the η^2 and partial η^2 statistics were calculated for each factor. Rather than using cut-off values for large effect sizes, the relative sizes of the η^2 statistics were compared within each outcome to determine which factor has the most influence on differences in the outcome measures. The findings of the simulation study are presented below.

RESULTS

Bias

The grand mean bias for each condition can be seen in Table 3. The largest bias (0.092) occurs within the 1-2-3 1/6 CAT 30-item MFI design, which appears larger relative to the other conditions. The smallest bias (.045) occurs within the 1-2-2 2/3 CAT 18 item random 5 MFI design. The smallest bias in the MST designs (.046) occurs within the 1-2-2 18 item design, while the largest bias in the MST designs (.069) occurs within the 1-2-2 30-item design. Table 3 showcases the variability in bias and shows that the 1-2-3 design tends to perform the worst in the ma-MST designs.

Table 3. Grand Mean Bias Across Conditions

CAT Proportion	MST Design	Random 1		Random 5	
		18 Item	30 Item	18 Item	30 Item
MST	1-2-2	0.046	0.069	---	---
MST	1-2-3	0.054	0.061	---	---
MST	1-3-3	0.057	0.063	---	---
1/6 CAT	1-2-2	0.051	0.076	0.056	0.076
1/6 CAT	1-2-3	0.088	0.092	0.077	0.081
1/6 CAT	1-3-3	0.046	0.062	0.056	0.065
1/3 CAT	1-2-2	0.047	0.075	0.050	0.071
1/3 CAT	1-2-3	0.068	0.076	0.065	0.079
1/3 CAT	1-3-3	0.060	0.069	0.058	0.068
2/3 CAT	1-2-2	0.045	0.055	0.045	0.056
2/3 CAT	1-2-3	0.049	0.059	0.047	0.059
2/3 CAT	1-3-3	0.049	0.058	0.047	0.059

The ANOVA results for grand bias indicated that most interaction terms and main effects were significant (see Table 4), and the four-way interaction term remained in the model. However, the factors with the highest η^2 and η_p^2 were the main effects of test length (η^2 =.091, η_p^2 =.115) and CAT Proportion (η^2 = .089, η_p^2 =.112); these each explained about 11% of the unexplained variance in the mean bias. Panel design and the interaction between panel design and CAT proportion, the factors with the next largest η^2 and η_p^2, explained about 5% of the unexplained variance in the mean bias each. The other main effects, two-way and three-way interactions, were either non-significant or explained a very small proportion of mean bias variance.

Table 4. ANOVA Results for Grand Mean Bias

Factor	df	SS	MS	*F* value	*p*	η^2	η^2_p
Panel Design	2	2064	1032	367.85	.000*	.041	.055
Length	1	4566	4566	1627.25	.000*	.091	.115
CAT Proportion	3	4625	1542	549.38	.000*	.089	.112
Random	1	1	1	0.22	.641	.000	.000
Panel Design: Length	2	530	265	94.38	.000*	.011	.015
Panel Design: CAT Proportion	6	2149	358	127.64	.000*	.034	.046
Length: CAT Proportion	3	108	36	12.86	.000*	.002	.003
Panel Design: Random	2	72	36	12.84	.000*	.001	.002
Length: Random	1	11	11	4.09	.043	.000	.000
CAT Proportion: Random	2	11	6	1.97	.140	.000	.000
Panel Design: Length: CAT Proportion	6	310	52	18.43	.000*	.005	.007
Panel Design: Length: Random	2	10	5	1.73	.177	.000	.000
Panel Design: CAT Proportion: Random	4	157	39	13.99	.000*	.003	.004
Length: CAT Proportion: Random	2	89	45	15.91	.000*	.002	.003
Panel Design: Length: CAT Proportion: Random	4	45	11	4.05	.003*	.001	.001
Residuals	12558	35239	3				
Total	12599	49987					

* Significant at the .05 level.

RMSE

The grand mean RMSE for each condition can be seen in Table 5. The largest RMSE (0.339) occurs within the 1-2-3 1/6 CAT 18 item random 5 MFI design, while the smallest RMSE (0.225) occurs within the 1-2-3 2/3 CAT 18 item random 5 MFI design. For the MST designs, the largest RMSE (0.327) occurs within the 1-2-2 18 item design, while the smallest RMSE (0.269) occurs within the 1-3-3 30 İtem design.

Table 5. Grand Mean RMSE Across Conditions

CAT Proportion	MST Design	Random 1		Random 5	
		18 Item	30 Item	18 Item	30 Item
MST	1-2-2	0.327	0.277	---	---
MST	1-2-3	0.318	0.310	---	---
MST	1-3-3	0.280	0.269	---	---
1-6 CAT	1-2-2	0.312	0.319	0.286	0.289
1-6 CAT	1-2-3	0.339	0.337	0.299	0.296
1-6 CAT	1-3-3	0.328	0.331	0.286	0.289
1-3 CAT	1-2-2	0.299	0.307	0.264	0.269
1-3 CAT	1-2-3	0.301	0.307	0.264	0.271
1-3 CAT	1-3-3	0.308	0.309	0.266	0.271
2-3 CAT	1-2-2	0.270	0.278	0.226	0.238
2-3 CAT	1-2-3	0.268	0.278	0.225	0.239
2-3 CAT	1-3-3	0.270	0.280	0.229	0.240

The ANOVA results for grand mean RMSE indicated that the four-way interaction between the factors was not significant, so Table 6 shows the ANOVA without this interaction term. Two factors dominated the variance explained RMSE -test length and CAT proportion- despite the significance of most of the interaction terms and all the main effects. The test length explained 36.7% of the total variance in RMSE, while the CAT proportion explained 42.6% of the total variance in RMSE. No other factors or interactions explained more than 5% of the total or unexplained variance in RMSE.

Table 6. ANOVA Results for Grand Mean RMSE

Factor	df	SS	MS	F value	*p*	η^2	η^2_p
Panel Design	2	181	91	58.59	.000*	.001	.009
Length	1	46689	46689	30199.84	.000*	.367	.706
CAT Proportion	3	57206	19069	12334.12	.000*	.426	.736
Random	1	924	924	597.834	.000*	.007	.045
Panel Design: Length	2	59	29	19.07	.000*	.000	.003
Panel Design: CAT Proportion	6	1792	299	193.22	.000*	.008	.048
Length: CAT Proportion	3	154	51	33.17	.000*	.001	.005
Panel Design: Random	2	6	3	1.82	.162	.000	.000
Length: Random	1	1	1	0.57	.451	.000	.000
CAT Proportion: Random	2	307	153	99.28	.000*	.002	.016
Panel Design: Length: CAT Proportion	6	296	49	31.94	.000*	.002	.011
Panel Design: Length: Random	2	1	0	0.32	.724	.000	.000
Panel Design: CAT Proportion: Random	4	46	12	7.45	.000*	.000	.002
Length: CAT Proportion: Random	2	20	10	6.36	.002*	.000	.001
Residuals	12562	19421	2				
Total	12599	127103					

* Significant at the .05 level.

Based on Table 6, test length was the most important factor on the RMSE and followed by CAT proportion. As the test length or CAT proportion increased, the amount of RMSE decreased.

Correlation

The grand mean correlation between the true and estimated theta values for each condition can be seen in Table 7. The smallest correlation (0.949) occurs within the 1-2-3 1/6 CAT 30 item random 1 MFI design, while the largest correlation (0.980) occurs within the 1-2-2 2/3 CAT 30 item random 5 MFI design. The MST design with the smallest correlation (0.950) was the 1-2-2 18 item design, while the largest correlation (0.971) occurred in the 1-3-3 30 item design.

Table 7. Grand Mean Correlation Across Conditions

CAT Proportion	MST Design	Random 1		Random 5	
		18 Item	30 Item	18 Item	30 Item
MST	1-2-2	.950	.970	---	---
MST	1-2-3	.957	.957	---	---
MST	1-3-3	.970	.971	---	---
1-6 CAT	1-2-2	.955	.955	.968	.967
1-6 CAT	1-2-3	.951	.949	.966	.966
1-6 CAT	1-3-3	.954	.952	.970	.970
1-3 CAT	1-2-2	.960	.958	.974	.972
1-3 CAT	1-2-3	.959	.958	.973	.972
1-3 CAT	1-3-3	.958	.958	.971	.972
2-3 CAT	1-2-2	.967	.966	.979	.980
2-3 CAT	1-2-3	.967	.964	.979	.978
2-3 CAT	1-3-3	.967	.964	.978	.976

Note. All correlations were significant at the alpha level of .05

The ANOVA results for grand correlation can be seen in Table 8. Like the grand mean RMSE, the four-way interaction between all factors was not significant and was removed from the ANOVA. Additionally, the same pattern of η^2 and η_p^2 was found: the highest values were found for the test length and CAT proportion, which explain 62.2% and 24.4% of the total variance, respectively. No other factor or interaction of factors explained greater than 5% of the variance in mean correlations.

Table 8. ANOVA Results for Grand Mean Correlation

Factor	df	SS	MS	F value	*p*	η^2	η^2_p
Panel Design	2	1	1	6.92	.001	.000	.001
Length	1	6211	6211	84496.47	.000	.622	.871
CAT Proportion	3	2622	874	11891.51	.000	.244	.725
Random	1	32	32	440.50	.000	.003	.034
Panel Design: Length	2	0	0	2.07	.126	.000	.000
Panel Design: CAT Proportion	6	85	14	193.83	.000	.003	.034
Length: CAT Proportion	3	60	20	271.37	.000	.004	.042
Panel Design: Random	2	1	0	4.36	.013	.000	.001
Length: Random	1	1	1	20.33	.000	.000	.002
CAT Proportion: Random	2	7	3	44.94	.000	.001	.007
Panel Design: Length: CAT Proportion	6	44	7	99.49	.000	.002	.022
Panel Design: Length: Random	2	0	0	0.04	.956	.000	.000
Panel Design: CAT Proportion: Random	4	1	0	3.32	.010	.000	.001
Length: CAT Proportion: Random	2	0	0	1.60	.202	.000	.000
Residuals	12562	923	0				
Total	12599	9988					

Based on Table 8, test length was the most important factor on the correlation and followed by CAT proportion. As the test length or CAT proportion increased, the size of correlation decreased.

DISCUSSION and CONCLUSION

This study aimed to determine how useful ma-MST, which follows the MST framework but utilizes an item-level adaptation routing module as in CAT, is in estimating theta as compared to standard MST designs. We hypothesized that ma-MST performs better than MST under the current simulation conditions according to grand mean bias, grand mean RMSE, and grand mean correlation. The results indicated that replacing the routing module of a certain length in a traditional MST with an equal-length item level adaptation routing module as in CAT results in similar levels of bias, lower levels of RMSE, and higher levels of correlation between true and estimated theta value. Including even more CAT items at the initial stage (the 2/3 CAT conditions) resulted in somewhat larger improvements in bias, RMSE, and correlation. The best-case scenarios for each outcome measure occurred within a 2/3 CAT condition, while the worst-case scenarios occurred within a 1/6 CAT condition. The most likely explanation for these results is that in 1/6 CAT condition, there were fewer items administered with item-level adaptation resulting in less accurate measures of ability in the routing stage, and the MFI item selection rule results in higher bias in the early stages of CAT (Chen, Ankenmann, & Chang, 2000).

The factors that were most important in determining the overall results were the test length and proportion of CAT items. Interestingly, tests with more items overall were associated with increased bias, although increasing the proportion of CAT items reduced the bias in every condition. This was counteracted with more items resulting in a smaller RMSE. This seeming contradiction is likely caused by a combination of the EAP estimator and by individuals at the boundaries of the module selection cutoffs (e.g., individuals with provisional ability estimates that caused the difference in the maximum module information to be small between the potential modules the individual could be routed to). The EAP estimator increases bias but decreases RMSE, particularly in more extreme values of ability (Kim, Moses, & Yoo, 2015). Improper routing of individuals is known to cause problems in MST, and the panel designs and module information functions in the simulation were not designed to prevent this from happening.

Unsurprisingly, we saw that the conditions with the highest proportion of CAT-based routing had the lowest levels of bias and RMSE as well as the highest correlations between the predicted and simulated theta values. However, since the ma-MST method provided good or better outcomes when the CAT routing panel was at least as large as a typical MST, the overall conclusion is that there is evidence to support the use of this design in circumstances that allow its use. For researchers and practitioners who wish to maintain many of the benefits of MST while improving its estimation efficiency, ma-MST is one method they should consider using.

While the study demonstrates the usefulness of ma-MST, it does so only for conditions that are similar to those in the simulation study. Another simulation with more varied conditions, such as different content balancing requirements, different unidimensional IRT models (e.g., 1PL or 2 PL), multidimensional IRT models, or estimation procedures, can further establish the usefulness of this approach, as well as a study comparing the designs with real data. Utilizing better panel designs which minimize the likelihood of misrouting or allow for misrouted individuals the chance to be re-routed into appropriate modules may provide more evidence of the efficacy of ma-MST over MST. Changes to the item and/or module selection method in ma-MST (e.g., by using a different information function) may also help improve the performance of the method as prior research has shown the choice of routing method can affect the efficacy of MST (Raborn, 2018). Another criticism in this study would be that the choice of some of the study conditions in the research design, especially for the ratio for the CAT proportion, is somewhat arbitrary. However, this study is an initial investigation of ma-MST approach. Finally, future research should investigate other ability estimation procedures such as maximum likelihood estimation as they may affect the relative efficiency of ma-MST when compared to MST.

As there have been other proposed combinations of CAT and MST in the literature such as Hybrid Computerized Adaptive Testing proposed by Wang et al. (2016), future research should include a comparison with these combinations as well as with full CAT tests. Investigating other simulation conditions that would serve to limit the limitations in this study would provide additional evidence for or against ma-MST in more circumstances.

REFERENCES

Armstrong, R. D., Jones, D. H., Li, X., & Wu, L. (1996). A study of a network-flow algorithm and a noncorrecting algorithm for test assembly. *Applied Psychological Measurement*, *20*(1), 89-98. doi: 10.1177/014662169602000108

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2008). Incorporating randomness in the Fisher information for improving item-exposure control in CATs. *British Journal of Mathematical and Statistical Psychology*, *61*(2), 493-513. doi: 10.1348/000711007X230937

Barton, M. A., & Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item-response model* (ETS RR-81-20). Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2333-8504.1981.tb01255.x

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* (pp. 17-20). Reading, MA: Addison-Wesley.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*(4), 431-444. doi: 10.1177/014662168200600405

Briethaupt, K., & Hare, D. R. (2007). Automated simultaneous assembly of multistage testlets for a high-stakes licensing examination. *Educational and Psychological Measurement, 67*(1), 5-20. doi: 10.1177/0013164406288162

Chen, S. Y., Ankenmann, R. D., & Chang, H. H. (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement*, *24*(3), 241-255. doi: 10.1177/01466210022031705

ILOG. (2006). *ILOG CPLEX 10.0* [User’s Manual]. Paris: ILOG SA. Retrieved from https://www.lix.polytechnique.fr/~liberti/teaching/xct/cplex/usrcplex.pdf

Kim, H., & Plake, B. S. (1993). *Monte Carlo simulation comparison of two-stage testing and computerized adaptive testing.* Atlanta, GA: National Council on Measurement in Education.

Kim, S., Moses, T., & Yoo, H. (2015). A comparison of IRT proficiency estimation methods under adaptive multistage testing. *Journal of Educational Measurement*, *52*(1), 70-79. doi: 10.1111/jedm.12063

Liu, O. L., Bridgeman, B., Gu, L., Xu, J., & Kong, N. (2015). Investigation of response changes in the GRE revised general test. *Educational and Psychological Measurement*, *75*(6), 1002-1020. doi: 10.1177/0013164415573988

Luecht, R. M., & Nungester, R. J. (2000). Computer-adaptive sequential testing. In W. J. van der Linden, & C. A. Glas, *Computerized adaptive testing: Theory and practice* (pp. 117-128). Netherlands: Springer.

Luecht, R. M., & Sireci, S. (2012). *A review of models for computer-based testing.* New York: The College Board. Retrieved from https://files.eric.ed.gov/fulltext/ED562580.pdf

Luecht, R. M., Brumfield, T., & Breithaupt, K. (2006). A testlet assembly design for adaptive multistage . *Applied Measurement in Education*, *19*(3), 189-202. doi: 10.1207/s15324818ame1903_2

Magis, D., Yan, D., & von Davier, A. (2017). *mstR: Procedures to generate patterns under multistage testing*. Retrieved from https://CRAN.R-project.org/package=mstR

Patsula, L. N. (1999). *A comparison of computerized adaptive testing and multistage testing* (Unpublished doctoral dissertation). University of Massachusett, Arherst, MA.

R Development Core Team . (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org

Raborn, A. W. (2018). *Package ‘caMST’*. Retrieved from https://cran.r-project.org/web/packages/caMST/caMST.pdf

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. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1087.9450&rep=rep1&type=pdf

Robin, F., Steffen, M., & Liang, L. (2016). The multistage test implementation of the GRE revised general test. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing* (pp. 363-380). Boca Raton, FL: Chapman and Hall/CRC.

Sari, H. I., & Huggins-Manley, A. C. (2017). Examining content control in adaptive tests: Computerized adaptive testing vs. computerized adaptive multistage testing. *Educational Sciences: Theory & Practice*, *17*(5), 1759-1781. doi: 10.12738/estp.2017.5.0484

Sari, H. I., Yahsi-Sari, H., & Huggins-Manley, A. C. (2016). Computer adaptive multistage testing: Practical issues, challenges and principles. *Journal of Measurement and Evaluation in Education and Psychology*, *7*(2), 388-406. Retrieved from https://dergipark.org.tr/en/download/article-file/270019

van der Linden, W. J. (2000). Constrained adaptive testing with shadow tests. In W. J. van der Linden, & G. A. W. Glas (Eds.), *Computer adaptive testing: Theory and practice* (pp. 27-52). Dordrecht: Kluwer Academic Publishers. doi: 10.1007/0-306-47531-6

Wang, S., Lin, H., Chang, H. H., & Douglas, J. (2016). Hybrid computerized adaptive testing: From group sequential design to fully sequential design. *Journal of Educational Measurement*, 53(1), 45-62. doi: 10.1111/jedm.12100

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*(4), 361-375. doi: 10.1111/j.1745-3984.1984.tb01040.x

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

Zenisky, A., Hambleton, R. K., & Luecht, R. M. (2010). Multistage testing: Issues, designs, and research. In W. J. van der Linden, & C. A. Glass (Eds.), *Elements of adaptive testing* (pp. 355-372). New York: Springer.

Appendix A: Annotated R Codes

An early version of the caMST package was used to perform the analyses in this simulation study. The analysis could be performed in v0.1.0 of the package (available on CRAN and GitHub; a developmental version is also available on the first author's GitHub repository); this is the version used for the brief demonstration here. This walkthrough assumes that you have a working R installation have installed the package with 'install_packages("caMST")' or 'devtools::install_github("AnthonyRaborn/caMST")', and for simplicity's sake only two conditions are shown: the CMT condition with the 1-3-3 panel design, equal-length routing, stage 2, and stage 3 modules, and 18 items, and the Ma-MST condition with the 1-3-3 panel design, 1/3 CAT routing module, 18 items, and MFI item selection in the routing stage.

This version of the package can only handle dichotomous IRT models and requires that four item parameters be specified for each of the items as in the four-parameter logistic model (4PL; Burton & Lord, 1981). That means that item parameters, as in most computer adaptive tests, are treated as fixed, known quantities and when using models other than the 4PL the equivalent item parameters still need to be specified. For example, if the item parameters being used come from the Rasch model, the discrimination and upper asymptote parameters for each item should be set equal to 1 and the guessing parameter for each item set equal to 0. As our simulation used 3PL items, the upper asymptote for all of our items was equal to 1, but the other three parameters varied as described in the text.

The main functions for this analysis were the *multistage_test* function, used for traditional MST formats, and the *mixed_adaptive_test* function, used for the ma-MST format. The data used in this study were simulated as explained above; item parameters were saved in a data frame with items on the rows and item parameters on the columns. To use the item parameter data frame with either of these functions, it should have the item parameters in the following format: item discriminations in column 1 named *a*, item difficulties in column 2 named *b*, the pseudo-guessing parameter in column 3 named *c*, and the upper asymptote in column 4 named *u*. Additionally, column 5 should be used for identifying the content area in which each item should be placed (if content balancing is needed) and is named *content_ID*. As of now, the item parameters must be formatted in this way for the functions to work.

From here, the *multistage_test* function will be used to demonstrate how we used the package functions for this study, then we will return to the *mixed_adaptive_test* function to highlight the differences in how the Ma-MST method is used.

The main function arguments for the *multistage_test* function are as follows:

- mst_item_bank: a matrix or data frame with the items formatted as above that contains all of the items that are used within this test. The rows of this data frame may be named to allow for the responses to be matched to the correct items automatically.
- modules: a matrix that relates the items in mst_item_bank to the modules in which they belong
- transition_matrix: a matrix that describes the possible modules individuals may be routed through
- response_matrix: a matrix or data frame of individuals' responses to the items in mst_item_bank, with persons on the rows and items on the columns. The item responses may be in the same order as in mst_item_bank: the first column of response_matrix should be the item in the first row of mst_item_bank. If not, the columns should share the same naming format as the rows of the mst_item_bank data frame to allow for the responses to be matched to the correct items automatically.
- n_stages: a numeric value indicating the number of stages in the test (e.g., the number of adaptation points plus one for the routing stage).
- test_length: a numeric value indicating the total number of items individuals will see.

Other options exist which allow for greater control over the way the item responses are analyzed; the function documentation goes into more detail.

For the 1-3-3 18-item CMT condition, the items we used are included with the package and can be called with the following commands:

```
## library(caMST)
## data(mst_only_items)
```

This will create the mst_only_items object in your global environment, which is a data frame with 42 rows (items) and 5 columns (item parameters). Using the head() function on this object shows the first six items and their paramaters (see Table A1).

These items were already placed in order in terms of the module they came it; that is, since each module has six items and there are seven modules across the three stages, the first six items are in the routing module, the second set of six items are in the first module at the second stage (the easy module), the third set of six items are in the second moduel at the second stage (the medium module), and so on. The item-module matrix for this data can be called into the environment with

```
## data(mst_only_matrix)
```

and is a 42 row (items) by 7 column (modules) matrix that looks like in Equation A1.

$$\begin{matrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{matrix} \tag{A1}$$

The next argument specifies the relationship between the modules (e.g., the lines in Figure 2). The 1-3-3 design we used in the simulation allows for individuals to move from one module in a stage to modules in the next stage that are the same difficulty or slightly more/less difficult, but does not allow for complete crossover. This means that a person routed to the stage 2 easy module may be placed into the stage 3 easy or medium difficulty modules but not the hard difficulty module. The transition matrix codifies this relationship using 0s to indicate that an individual in the row's module cannot be placed in the column's module and 1s to indicate that they could be placed from the row's module to the column's module. The matrix for this condition is called with

```
## data(example_transition_matrix)
```

and looks like in Equation A2.

$$\begin{matrix} 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix} \tag{A2}$$

The transition matrix should always be a square matrix with row and column sizes equal to the number of modules in the data, and the rows for the final stage modules should always be filled with 0 because there is no transition after the test is complete!

The response_matrix is simply the matrix or data frame of person responses. The package functions will try to use the column names of the response_matrix and the row names of the mst_item_bank data frame to extract the responses relevant to the current condition. Since caMST can only handle binary items, the responses should be all 0s, 1s, and NAs. An example set of responses from this simulation can be called with data(example_responses), which populates the current R environment with a 5 row (individuals) by 600 column (items) data frame of responses.

With these objects, we can use the multistage_test function to analyze the item responses as a CMT with the following code:

```
## multistage_test(mst_item_bank = mst_only_items, modules = mst_only_matrix,
## transition_matrix = example_transition_matrix,
## response_matrix = example_responses, n_stages = 3, test_length = 18)
```

The function will output a list of the results, which includes two different estimates of the individuals' abilities, the standard error of measurement for each individual, a matrix of the final items seen by all individuals, a matrix of the final modules seen by all individuals, and a matrix of the responses the individuals made to the items that they saw.

By changing the mst_only_items, mst_only_matrix, example_transition_matrix, n_stages, and test_length, each of the conditions ran in the simulation can be tested. In addition, since we know the true theta values used in the simulation, the bias, RMSE, conditional bias, conditional RMSE, conditional SEM, and correlation between true and estimated values are easily calculated with functions that take the true and estimated theta values as arguments. If the above results were saved as an object called *CMT_results*, calling *CMT_results$final.theta.estimate.mstR* produces these estimates and could be used for one of the functions. For example, assuming the true theta values are saved as a numerical vector called *example_thetas*, you could run *cor(example_thetas, CMT_results$final.theta)* to estimate the correlation between the values the responses were simulated from and the estimates from the *multistage_test* function.

The Ma-MST conditions were run with the *mixed_adaptive_test* function, which follows the same principles as the CMT function. The major difference is that the *mixed_adaptive_test* function requires two item banks: one for the first stage with item-level adaptation (i.e., for the CAT-style routing module), and another for the second and third stages (i.e., the CMT-style stages). Additionally, the function allows for some control over the way the CAT adaptation in the first stage occurs.

The arguments specific to the CAT routing module are:

- cat_item_bank: the item bank formatted as described in the “multistage_test” function
- item_method: the method for choosing items in the first stage; defaults to “MFI” (Maximum Fisher Information), which we used in our simulation
- cat_length: how many items are seen in the first stage
- cbControl: a list used for content balancing (not used in this study)
- cbGroup: a factor vector used for content balancing (not used in this study)
- randomesque: an integer value. The item_method ranks items from best to worst; using MFI and randomesque=1, the most informative item based on the Fisher information and the current response pattern is chosen, while using MFI and randomesque=5 will randomly select one item from the five most informative items based on the Fisher information and the current response pattern.

The arguments specific to the CMT modules are:

- mst_item_bank: the item bank formatted as described in the “multistage_test” function
- transition_matrix: a matrix that describes the possible modules individuals may be routed through

When comparing the two functions, it is easy to see that the addition of the CAT items is the only real change in the function arguments. The following code calls the new objects (one for the routing module items, another for the second and third stage items) and runs the Ma-MST 1-3-3 design 18 items 1/3 CAT MFI condition:

```
## data(cat_items); data(mst_items)
## mixed_adaptive_test(cat_item_bank = cat _items, cat_length = 6, item_method = “MFI”,
## randomesque = 1,
## mst_item_bank = mst_items, modules = mst_only_matrix,
## transition_matrix = example_transition_matrix,
## response_matrix = example_responses, n_stages = 3)
```

The results for this function contain the same information as the previous function, but are in a list format where each individual's entire results are saved as one element of that list. This helps when keeping track of the items each individual saw: the function keeps a track of the item parameters of each item seen by each individual and provides the individualized test bank as a part of the output.

Since the results of this function are in a list, it takes a little more effort to use them to test how well the method performs in terms of person parameter recovery. The easiest way to do this is by using the *getElement* function, which takes an object and the name of the element you wish to extract, within the *sapply* function, which applies one function to each element of another object. Putting these together will extract the information into a vector, similar to what the *multistage_test* function outputs automatically. If the output of the *mixed_adaptive_test* function was saved as *results*, then running

```
## sapply(results, getElement, "final.theta.estimate.mstR", simplify = T)
```

will output a vector of the final estimated theta values. This can then be used as explained above to investigate the efficiency of the Mca-MST method under the specific conditions used.

By modifying the various function arguments and the objects used in the functions, this study could be replicated or even expanded relatively easily. The package documentation includes other examples, as well as a function for performing fully CAT-formatted tests. The readme file and GitHub website provide somewhat more in-depth examples with visuals on the input and output data.

Table A1. The First Six Items for the CMT Condition

Item	*a*	*b*	*c*	*u*	content_ID
Item7	1.534	0.216	0.163	1.000	1
Item24	1.458	-0.136	0.070	1.000	1
Item165	1.696	-0.189	0.190	1.000	2
Item187	1.735	-0.024	0.097	1.000	2
Item303	1.410	0.243	0.068	1.000	3
Item458	1.446	-0.475	0.277	1.000	4

Note: a is the item discrimination, b is the item difficulty, c is the item pseudo-guessing parameter, and u is the upper asymptote of the item function.

Appendix B: Supplementary Tables and Figures

Table B1. Content Distributions in the Modules in the 1/6 CAT Conditions Across the Different Designs

	18-item						30-item					
Design	S2E	S2M	S2H	S3E	S3M	S3H	S2E	S2M	S2H	S3E	S3M	S3H
1-2-2	C1:2	-	C1:2	C1:1	-	C1:1	C1:4	-	C1:4	C1:3	-	C1:3
	C2:2		C2:2	C2:2		C2:2	C2:3		C2:3	C2:3		C2:3
	C3:2		C3:2	C3:2		C3:2	C3:3		C3:3	C3:3		C3:3
	C4:2		C4:2	C4:2		C4:2	C4:3		C4:3	C4:3		C4:3
Total	8	-	8	7	-	7	13		13	12	-	12
1-2-3	C1:2	-	C1:2	C1:1	C1:1	C1:1	C1:4	-	C1:4	C1:3	C1:3	C1:3
	C2:2		C2:2	C2:2	C2:2	C2:2	C2:3		C2:3	C2:3	C2:3	C2:3
	C3:2		C3:2	C3:2	C3:2	C3:2	C3:3		C3:3	C3:3	C3:3	C3:3
	C4:2		C4:2	C4:2	C4:2	C4:2	C4:3		C4:3	C4:3	C4:3	C4:3
Total	8		8	7	7	7	13		13	12	12	12
1-3-3	C1:2	C1:2	C1:2	C1:1	C1:1	C1:1	C1:4	C1:4	C1:4	C1:4	C1:4	C1:4
	C2:2	C2:2	C2:2	C2:2	C2:2	C2:2	C2:3	C2:3	C2:3	C2:3	C2:3	C2:3
	C3:2	C3:2	C3:2	C3:2	C3:2	C3:2	C3:3	C3:3	C3:3	C3:3	C3:3	C3:3
	C4:2	C4:2	C4:2	C4:2	C4:2	C4:2	C4:3	C4:3	C4:3	C4:3	C4:3	C4:3
Total	8	8	8	7	7	7	13	13	13	13	13	13

S = Stage, E = Easy, M = Medium, H = Hard module

Table B2. Content Distributions in the Modules in the 1/3 CAT Conditions Across the Different Designs

	18-item						30-item					
Design	S2E	S2M	S2H	S3E	S3M	S3H	S2E	S2M	S2H	S3E	S3M	S3H
1-2-2	C1:1	-	C1:1	C1:1	-	C1:1	C1:2	-	C1:2	C1:3	-	C1:3
	C2:1		C2:1	C2:1		C2:1	C2:2		C2:2	C2:3		C2:3
	C3:2		C3:2	C3:2		C3:2	C3:3		C3:3	C3:1		C3:1
	C4:2		C4:2	C4:2		C4:2	C4:3		C4:3	C4:3		C4:3
Total	6		6	6		6	10		10	10		10
1-2-3	C1:1	-	C1:1	C1:1	C1:1	C1:1	C1:2	-	C1:2	C1:3	C1:3	C1:3
	C2:1		C2:1	C2:1	C2:1	C2:1	C2:2		C2:2	C2:3	C2:3	C2:3
	C3:2		C3:2	C3:2	C3:2	C3:2	C3:3		C3:3	C3:1	C3:1	C3:1
	C4:2		C4:2	C4:2	C4:2	C4:2	C4:3		C4:3	C4:3	C4:3	C4:3
Total	6		6	6	6	6	10		10	10	10	10
1-3-3	C1:1	C1:1	C1:1	C1:1	C1:1	C1:1	C1:2	C1:2	C1:2	C1:3	C1:3	C1:3
	C2:1	C2:1	C2:1	C2:1	C2:1	C2:1	C2:2	C2:2	C2:2	C2:3	C2:3	C2:3
	C3:2	C3:2	C3:2	C3:2	C3:2	C3:2	C3:3	C3:3	C3:3	C3:1	C3:1	C3:1
	C4:2	C4:2	C4:2	C4:2	C4:2	C4:2	C4:3	C4:3	C4:3	C4:3	C4:3	C4:3
Total	6	6	6	6	6	6	10	10	10	10	10	10

S = Stage, E = Easy, M = Medium, H = Hard module

Table B3. Content Distributions in the Modules in the 2/3 CAT Conditions Across the Different Designs

	18-item						30-item					
Design	S2E	S2M	S2H	S3E	S3M	S3H	S2E	S2M	S2H	S3E	S3M	S3H
1-2-2	C1:1	-	C1:1	C1:0	-	C1:0	C1:1	-	C1:1	C1:1	-	C1:1
	C2:0		C2:0	C2:1		C2:1	C2:1		C2:1	C2:1		C2:1
	C3:1		C3:1	C3:1		C3:1	C3:1		C3:1	C3:2		C3:2
	C4:1		C4:1	C4:1		C4:1	C4:2		C4:2	C4:1		C4:1
Total	3		3	3		3	5		5	5		5
1-2-3	C1:1	-	C1:1	C1:0	C1:0	C1:0	C1:1	-	C1:1	C1:1	C1:1	C1:1
	C2:0		C2:0	C2:1	C2:1	C2:1	C2:1		C2:1	C2:1	C2:1	C2:1
	C3:1		C3:1	C3:1	C3:1	C3:1	C3:1		C3:1	C3:2	C3:2	C3:2
	C4:1		C4:1	C4:1	C4:1	C4:1	C4:2		C4:2	C4:1	C4:1	C4:1
Total	3		3	3	3	3	5		5	5	5	5
1-3-3	C1:1	C1:1	C1:1	C1:0	C1:0	C1:0	C1:1	C1:1	C1:1	C1:1	C1:1	C1:1
	C2:0	C2:0	C2:0	C2:1	C2:1	C2:1	C2:1	C2:1	C2:1	C2:1	C2:1	C2:1
	C3:1	C3:1	C3:1	C3:1	C3:1	C3:1	C3:1	C3:1	C3:1	C3:2	C3:2	C3:2
	C4:1	C4:1	C4:1	C4:1	C4:1	C4:1	C4:2	C4:2	C4:2	C4:1	C4:1	C4:1
Total	3	3	3	3	3	3	5	5	5	5	5	5

S = Stage, E = Easy, M = Medium, H = Hard module

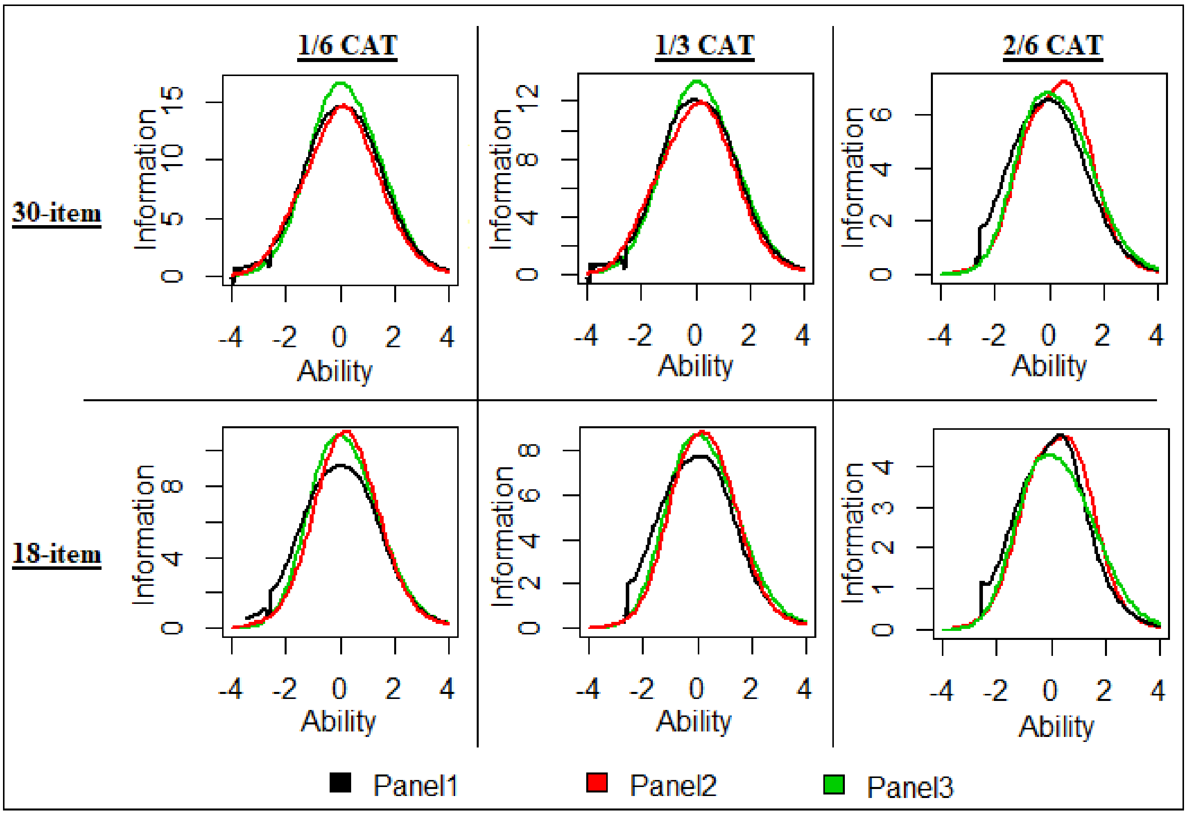

Figure B1. Plots for the Three Panels Under 1-2-2 ma-MST Design Across the 30-Item (Upper Three) and 18-Item (Down Three) and 1/6 CAT (Left two), 1/3 CAT (Middle two) and 2/6 CAT (Right two) Conditions