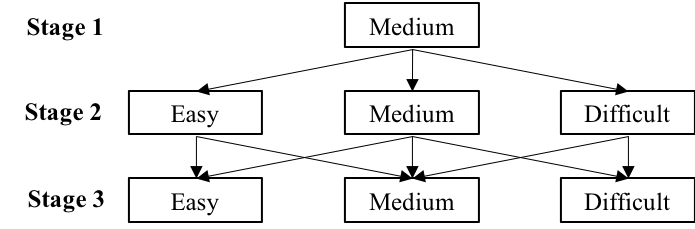
**Effect of Routing Errors on the   
Psychometric Properties of Multistage Tests**

Two-stage tests, the simplest implementation of a multistage test (MST), were the first type of adaptive tests proposed to replace conventional paper-and-pencil tests (e.g., Angoff & Huddleston, 1958; Cleary, Linn, & Rock, 1969; Linn, Rock, & Cleary, 1969, Weiss & Betz, 1973). In a two-stage test, examinees begin with a short fixed-form conventional test. That test is scored, and based on the score the examinee is routed to a second stage test of varying difficulty. Examinees with low scores on the first-stage test are routed or branched to an easy set of items, those with scores near the center of the score distribution are routed to a test of medium difficulty, and those with high scores receive a test with more difficult items. An examinee’s score on the test is based on their responses to both sets of items that they have answered.

Early research (Angoff & Huddleston, 1958: Betz & Weiss, 1973, 1974; Cleary et al., 1969, Linn et al., 1969; Larkin & Weiss, 1975; Lord, 1971b) suggested that two-stage tests were a viable type of adaptive test, resulting in psychometric improvements over fixed-form conventional tests. However, the early research on two-stage tests waned as alternative adaptive test designs that implemented more than minimal levels of adaptation were proposed and evaluated (e,g., Hornke, 1979; Lord, 1971a, 1971c; Vale & Weiss, 1975; Waters, 1977; Weiss, 1974). By the late 1970s most adaptive testing research and development focused on fully adaptive computerized adaptive tests (CATs) utilizing item response theory (IRT), in which a new item is selected from a pre-calibrated item bank for each examinee after each item is answered.

Until the mid-1990s there was very little research and almost no applications of two-stage tests or MSTs. However, MSTs recently have been proposed as an alternative to CATs and have been implemented using IRT methods for test design and ability estimation (Yan, von Davier, & Lewis, 2014). Figure 1 shows an example of a three-stage MST. As in a two-stage test, in Stage 1, all examinees begin the test with a set of items with medium difficulty. Each examinee is then routed to one of the three item sets in Stage 2 that is most informative given their IRT ability (*θ*) estimate at the end of Stage 1. After Stage 2, *θ* is re-estimated and another routing decision is made. Using the terminology of Luecht and Nungester (1998), the sets of items at each stage are called *modules*. MST designs are often expressed numerically with a shorthand method (Yan et al., 2014). For example, the MST in Figure 1 can be referred to as a 1-3-3 design, meaning that there is one module in Stage 1, three modules in Stage 2, and three modules in Stage 3. A collection of modules in a test form is called a *panel*. It is common practice to build multiple panels and randomly assign one of them to examinees (Yan et al., 2014)

**Figure 1. A 1-3-3 MST design**



**Routing Errors**

Early evaluations of two-stage tests (and by extension, MSTs) identified routing errors as a potential problem in MSTs. A routing error is said to occur if an examinee is branched or routed to a module that is not well matched with his/her true *θ* level. This can occur when examinees with low *θ* have a series of lucky guesses, or when examinees with high *θ* have a poor start for some reason. Routing errors might also occur simply as the result of measurement error due to a non-optimal item bank, items with low information, mismatch between the examinee’s *θ* level and item information, or other factors that contribute to measurement error.

Lord (1974) stated “…optimal assignment of examinees to levels on a multilevel test can never be perfectly achieved without knowing examinee ability. Thus, some examinees are always misassigned” (p, 7)—an observation independently echoed by Betz and Weiss (1973). Early empirical research on routing errors observed percentages of routing errors that varied from about 1% (Larkin & Weiss, 1975) and 5% (Betz & Weiss, 1973), to 20% (Angoff & Huddleston, 1958), to as much as 40% (Cleary et al. 1969), depending on the criterion used to identify misrouting.

Curiously, however, there appears to have been no research on misrouting for the 30-year period beginning in 1974 until Kim and Moses (2014) examined the potential impact of routing errors on the measurement performance of two-stage tests. They simulated two 1-3 MSTs under different conditions. In the small-difference condition, the three modules of the second stage overlapped in difficulty, whereas in the large-difference condition, they did not. In each condition, all three possible paths were administered to a sample of simulees. The total number-correct scores from each of the three paths were obtained and equated using item response theory (IRT) true-score equating. They found that the score differences associated with different paths were negligible and the results from the two conditions were almost indistinguishable. However, for reasons they did not explain, a constant discrimination parameter was used in the process of equating, even though the discrimination parameters of items they simulated were not constant. In addition, their results are limited in generalizability due to the use of number-correct scores to estimate examinee ability and implement routing.

Luo and Kim (2018) reported the only other study to explicitly address misrouting. Their Monte Carlo simulation, which made strict assumptions about the distributions of routing errors, compared MSTs with and without a routing error control procedure they proposed. They concluded that MSTs with routing error control showed lower root mean squared errors (RMSEs) and that it showed misrouting played a role in reducing measurement precision. However, in six of the nine conditions they examined, the differences in RMSE were smaller than 0.02, and the maximum difference in all nine conditions was only 0.05. Thus, it could be that the impact of routing error was so small that there was not much room to improve, or that the routing error control procedure they proposed was not effective enough. Either way, they did not provide a direct measure of the impact of routing errors, although their graphic results suggested that there were substantial numbers of routing errors in their results. Moreover, they did not report the results conditional on *θ*, so it was not clear which *θ* region was most affected by routing errors.

Finally, based on live testing with two-stage tests used in an NAEP study (Oranje, Mazzeo, Xu, & Kulick, 2014), the actual routing from the first-stage test was compared with the routing that would have occurred based on the examinee’s *θ* estimates from all items administered. They concluded that “The routing was quite accurate” because 85.9% of the routing decisions would have been the same, and when corrected for “measurement error” the agreement reduced to 81.5%, or a misrouting rate (aggregated across *θ* levels) of about 20%. It is clear from the extremely limited amount of research available on misrouting in MSTs that current implementations of MSTs have not carefully considered the impact of routing errors on the measurements they obtain from MSTs.

## Comparison of CATs and MSTs

Proponents of MSTs suggest that MSTs offer the potential to overcome certain shortcomings of CATs. First, since MSTs are pre-assembled, instead of being assembled on-the-fly like CATs, MSTs can be reviewed by subject matter experts before they are deployed, and hence are assumed to allow better control over content balancing (Yan et al., 2014). Second, item exposure control can easily be implemented by a random selection of panels, without the need for complex algorithms (Yan et al., 2014). Third, because CATs administer different subsets of items to different examinees, they typically result in a sparse item response matrix, which makes statistical analyses, such as differential item functioning and item recalibration, difficult. This problem is mitigated in MSTs as examinees are restricted to answer pre-defined sets of items. Hence, the sparse matrix from MSTs are block sparse, meaning that responses occur in regular patterns, and data analyses are therefore more tractable (Mead, 2006). Lastly, examinees are allowed to skip items in an MST, as well as review and change answers within a stage, whereas a typical CAT allows items to be skipped, but does not allow item review because of the dynamic nature of the CAT algorithm.

Since the primary purpose of testing is to measure individual differences, whether MSTs are a viable alternative to CATs should not be evaluated merely based on practical concerns but also on the measurement performance of the two approaches. This naturally leads to the question of which kind of assessment design performs better in terms of accuracy and precision of *θ* estimates. Although there are studies comparing the performance of CATs and MSTs, most of them (Hambleton & Xing, 2006; Xing & Hambleton, 2004; Zheng, Nozawa, Gao, & Chang, 2012) considered only situations where the purpose of the tests was to classify examinees into groups (e.g., making pass-fail decisions). However, MSTs have been adopted by several major testing programs to measure individual differences: for example, the Graduate Record Examination and the Massachusetts Adult Proficiency Tests (Yan et al., 2014). It is possible that a test performs well when making accurate classification decisions, but not when the goal of the test is to precisely measure individual differences.

Of the studies using MSTs that did focus on measurement precision, many were based on testlets in which a set of items were based on a common stimulus or scenario such as reading comprehension items (Davis & Dodd, 2003; Keng, 2008; Rotou, Patsula, Steffen, & Rizavi, 2007; Schnipke & Reese, 1999). Because of the dependency between items within a testlet, the findings derived from these studies might not necessarily be generalizable to tests that are not testlet-based.

Among the studies that compared item-level CATs and MSTs, their design did not permit meaningful comparisons of the two testing approaches. For example, Wang (2017) used different item banks for the CAT and the MST, which led to noncomparable results. In the simulation study by Kim & Plake (1993), the modified one-parameter logistic (1PL) model was used to calibrate the items and generate item responses. In the modified 1PL model, discrimination and pseudo-guessing parameters were constant for all items, resulting in findings that would not be likely to generalize to real testing situations in which item discrimination and guessing parameters are unlikely to be constant. Furthermore, no studies have compared MSTs with CATs in the presence of an evaluation of the impact of routing errors in the MSTs.

## MST Design Factors

Given the large range of possible test designs for MSTs, an optimal design should be used to evaluate the magnitude and effects of misrouting. However, the MSTs used in some studies (Wang, 2017; Zheng & Chang, 2015) appear to be sub-optimally configured and the effects of some MST design factors have not yet been thoroughly investigated.

The 1-3-3 design is a very popular test structure used in MST studies. However, some studies (Wang, 2017; Zheng & Chang, 2015) used the 1-3-3 design with difficulties of the three modules in the second and third stages both anchored at −1, 0 and 1. Thus, the third stage did not provide the potential to improve the *θ* estimates obtained at the second stage for examinees with extreme abilities, because items in the third stage were just as (un)informative as the items in the second stage.

Another design factor to consider is the item allocation. One rationale is to populate the later stages with more items so that examinees can receive more items that are closely matched to their ability. Another rationale is that more items should be placed in the early stage so that examinees can be routed to a module that is more suitable to their levels. Zheng et al. (2012) examined this problem and concluded that the two rationales made no difference. However, in their study, the longest stage had only three more items than the shortest stage. It is very possible that the range was too small to exhibit any effect. Moreover, their study focused on classification accuracy. Chen (2010) and Macken-Ruiz (2008) both found that the performance of MSTs with an increasing number of items per stage was better, but their studies were based on polytomous items.

The third factor is the assembly priority. In a forward assembled MST, modules in the early stages are built up prior to those in later stages. In a backward assembled MST, modules in the later stages are constructed first. Zheng et al. (2012) found that the backward assembly method produced higher classification accuracy, while Wang (2017) found that both designs produced similar mean biases and mean squared errors. Unfortunately, the study by Zheng et al. (2012) focused on classification accuracy, and the results by Wang (2017) were aggregated across all *θ* levels, so potential differences between *θ* levels might be missed.

## Purposes of the Study

The objectives of this study were threefold: (1) to examine how routing errors affect MST performance; (2) to investigate how certain MST design factors—test structure, item allocation, and assembly priority—affect the measurement performance of MSTs in the presence of routing errors; and (3) to compare the measurement precision and efficiency of *θ* estimation produced by item-level CATs and MSTs designed for measuring individual differences, with and without routing errors in the MSTs.

# Method

## Overview

Three MST design factors (test structure, item allocation and assembly priority) were manipulated. All MST designs were assembled from a “master” item bank. The items that were actually selected to be used for an MST will be referred to as an *operational bank*. Each MST design was matched with a CAT that used the same operational bank as the MST, so that the CAT and the MST would be comparable within a condition.

This study used Monte Carlo simulation methods. For each simulee in each condition, responses were simulated for all items in the operational bank, then both CATs and MSTs were applied to estimate *θ* based on that same set of item response. The objective of this procedure was to determine how the measurement performance would change, had a simulee taken a CAT instead of an MST, and vice-versa.

## Simulee Population

A total of 6,500 simulees were generated with 500 simulees each at *θ* levels ranging from 3 to 3 in increments of 0.5. A uniform distribution was used so that the precision of *θ* estimates and other dependent variables could be evaluated across the entire *θ* range, compared to a standard normal distribution that some other similar studies used (Keng, 2008; Wang, 2017; Zheng & Chang, 2015) resulting in those studies being unable to evaluate performance conditional on *θ*.

## Multistage Tests

The overall test length was fixed at 42 items. Since it is a common practice in implementations of MST to assemble parallel panels (Yan et al., 2014), the present study constructed five panels to make the simulation more realistic.

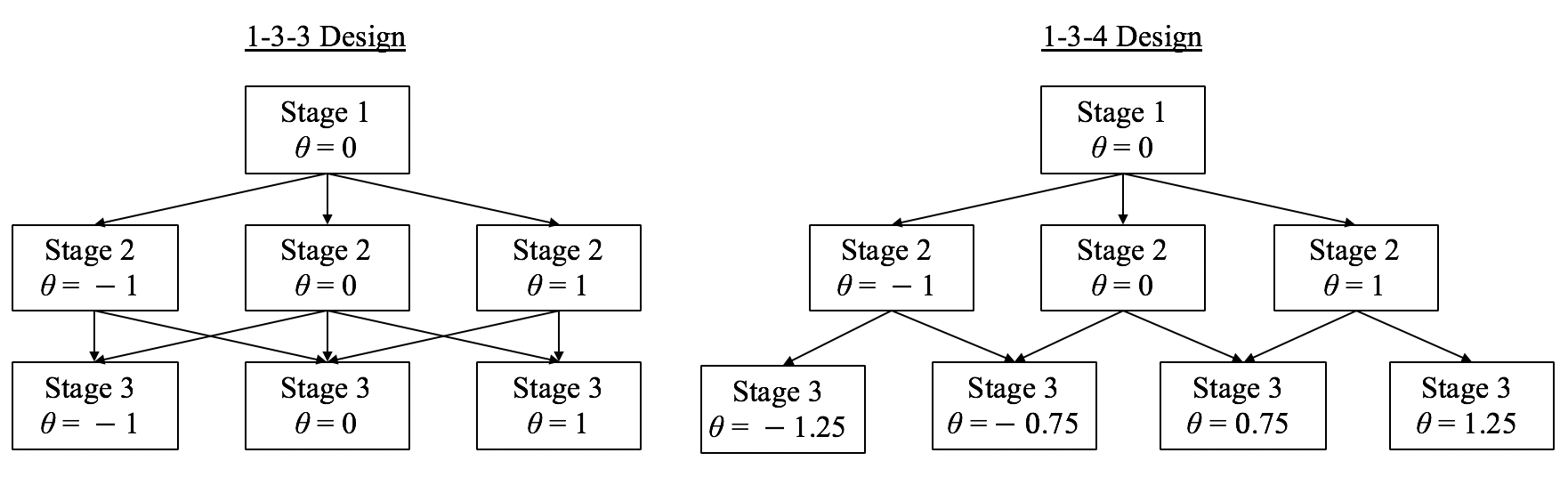
Conditions. Three MST design factors were manipulated to test which MST design yielded the best performance and to compare MSTs with CATs. This resulted in a total of 2 × 3 × 2 = 12 MST conditions as shown in Table 1.

**Table 1. Summary of MST designs**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Number of items | | |
| Test structure | Assembly priority | Stage 1 | Stage 2 | Stage 3 |
| 1-3-3 | Forward | 7 | 14 | 21 |
|  |  | 14 | 14 | 14 |
|  |  | 21 | 14 | 7 |
|  | Backward | 7 | 14 | 21 |
|  |  | 14 | 14 | 14 |
|  |  | 21 | 14 | 7 |
| 1-3-4 | Forward | 7 | 14 | 21 |
|  |  | 14 | 14 | 14 |
|  |  | 21 | 14 | 7 |
|  | Backward | 7 | 14 | 21 |
|  |  | 14 | 14 | 14 |
|  |  | 21 | 14 | 7 |

Test structure. Two test structures (1-3-3 and 1-3-4) were compared. For the 1-3-3 design, the present study followed the practice in Wang (2017), and Zheng and Chang (2015), where the second and third stages had the same set of difficulty anchors. For the 1-3-4 design, the *θ* anchors of the last stage were chosen to be the same as the design in Schnipke & Reese (1999). The *θ*s at which the module information was maximized are listed in Figure 2. Note that some pathways were restricted so that simulees were not allowed to move to a module in the next stage that had a difference of more than one level of difficulty as compared to the module in the current stage. This was to prevent a drastic change in *θ* estimates, because this would indicate non-model-fitting behavior and would be flagged as aberrant in practice (Chen, 2010; Jodoin, Zenisky, & Hambleton, 2006; Luecht, Brumfield, & Breithaupt, 2006).

Item allocation. Three levels of item allocation were evaluated. The increasing number of items per stage condition assigned items as [1/6 (7 items), 1/3 (14 items), 1/2 (21 items)], the decreasing number of items per stage condition assigned items as [1/2, 1/3, 1/6], and as a control, a condition that had an equal number of items per stage [1/3, 1/3, 1/3] was also used. These proportions were used in Patsula (1999). It ensured that the number of items were equally spaced across stages.

***Figure 2.* 1-3-3 and 1-3-4 MST designs** **

Assembly priority. The third factor was the assembly priority (forward and backward).

### Item Bank. A total of 1,500 items were generated using the 3-parameter logistic IRT model, where the probability of answering item *i* correctly for examinee *j* is defined as

where , and are the discrimination, difficulty and pseudo-guessing parameters respectively, is the ability level of examinee *j*, and D = 1.7 is used to scale the *ai* parameters from a logistic metric to the normal metric. Table 2 presents the descriptive statistics for the item parameters. As Wang (2013) has noted, the item bank size should be set as 1.5 times the number of items required. In this study, the largest number of items required for an MST design was 5 panels × [(7 items + (14 items × 3 modules) + (21 items × 4 modules)] = 560 items, so 1,500 items were determined to be more than sufficient.

Test Assembly. The bottom-up approach (Yan et al., 2014) was employed to achieve parallelism across panels. That is, for each module, five parallel forms were assembled. The

**Table 2. Descriptive statistics for item parameters**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Parameters | Mean | SD | Minimum | Maximum | Distribution |
| *a* | 1.03 | 0.27 | 0.41 | 2.20 | *ln*N(0.75, 0.25) |
| *b* | -0.01 | 1.03 | -3.25 | 3.81 | N(0, 1) |
| *c* | -0.20 | 0.03 | 0.15 | 0.25 | *Unif*(0.1, 0.2) |

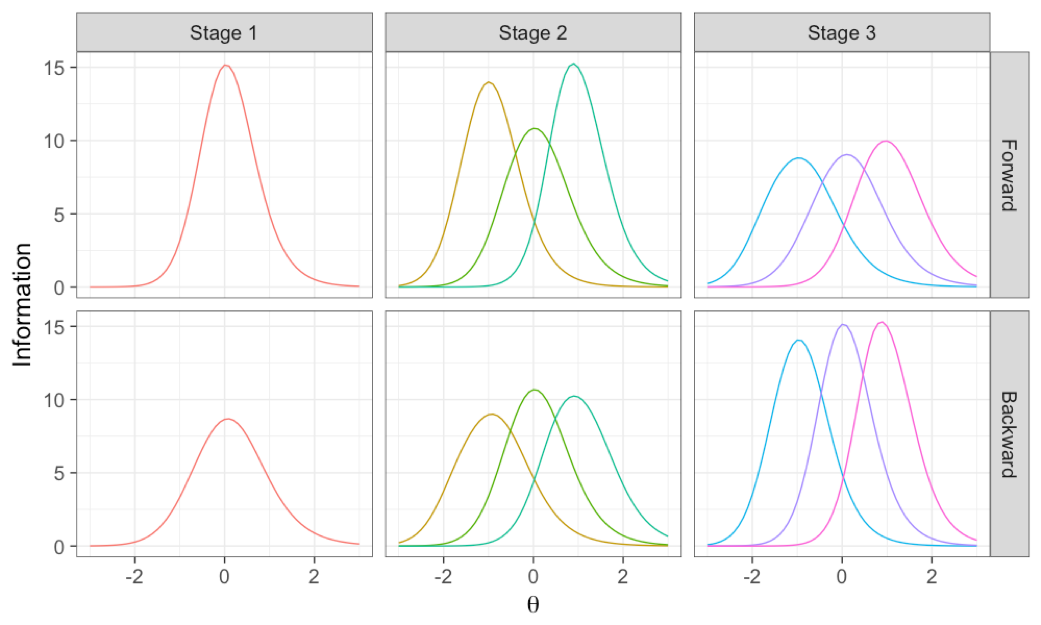
bottom-up approach was easy to implement because when the alternative forms of each module are parallel, corresponding pathways in the resulting panels will automatically be parallel. For each module, items with the most information at the corresponding *θ* anchor were selected. The information for item *i* is defined as

The order in which the stages received items depended on whether the design was forward assembled or backward assembled. Within a stage, every combination of module and panel had an equal probability to be chosen to select items. Figure 3 shows the module information functions (MIF) of the 1-3-3 and 1-3-4 MSTs with equal numbers of items per stage respectively, averaged over five panels. The MIFs of other conditions have been omitted for brevity. Figure 4 displays the panel information functions for each of the MST panels in each condition.

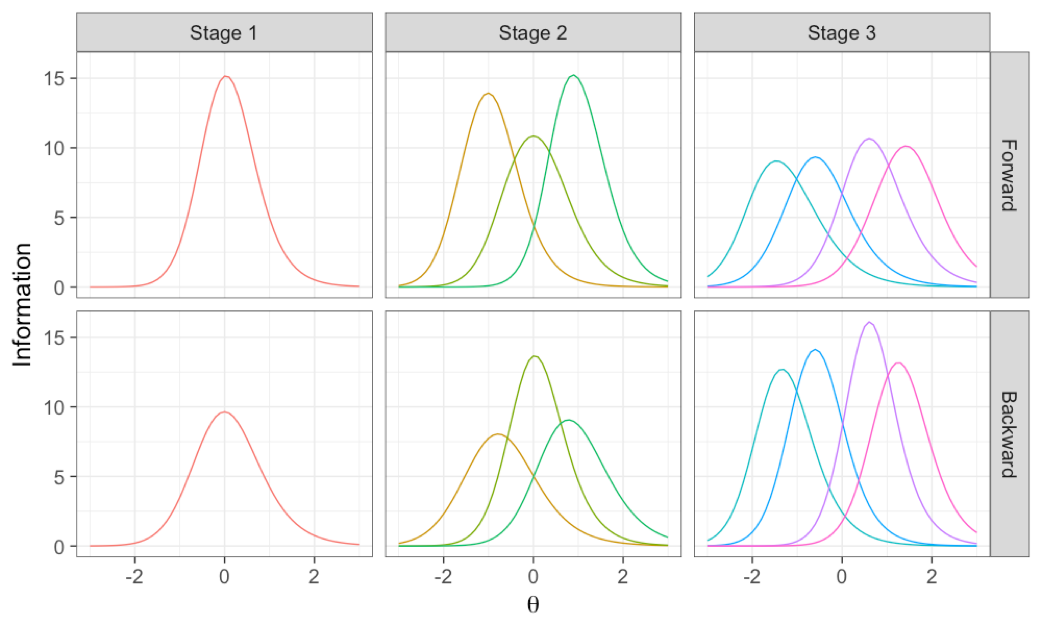
Test Administration. The MST administration was simulated using the mstR package (Magis, Yan, & von Davier, 2018) in R. Simulees were randomly assigned to one of the five panels and were routed to the next stage module that provided the most information at the current estimate (using all the items up to that point). Maximum likelihood estimation (MLE) was used to estimate θ. The range of θ estimates was (−3.5, 3.5), which was set to be larger than the range of true θ levels to minimize any floor or ceiling effect. The θ estimates were set to the upper bound value if the derivatives of the log-likelihood function were positive at both θ = −3.5 and θ = 3.5. On the other hand, the θ estimates were set to the lower bound value if both derivatives were negative.

**Figure 3. Module information functions in 1-3-3 and 1-3-4 MST designs, forward and backward assembled, with equal number of items per stage, averaged across five panels**

1. **1-3-3 MST**



1. **1-3-4 MST**

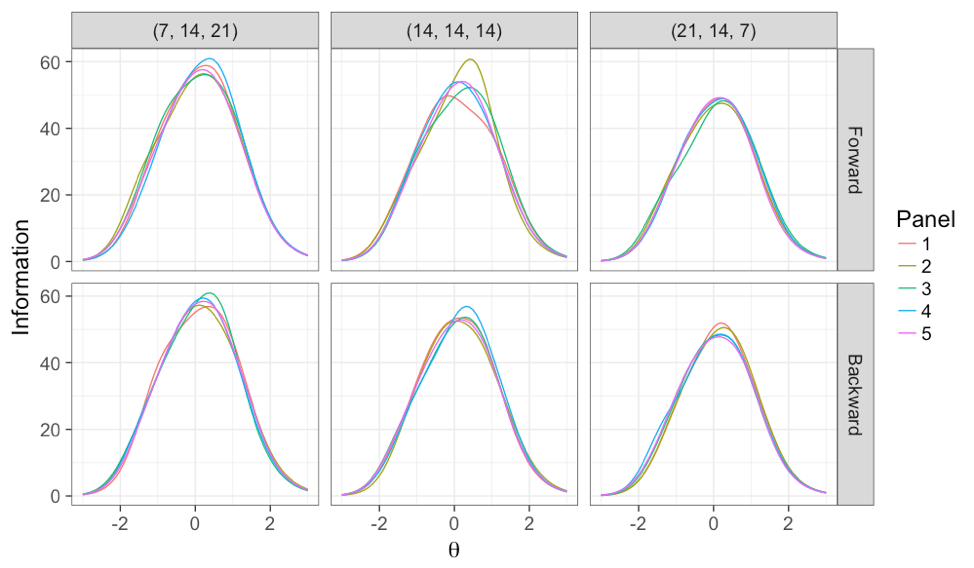


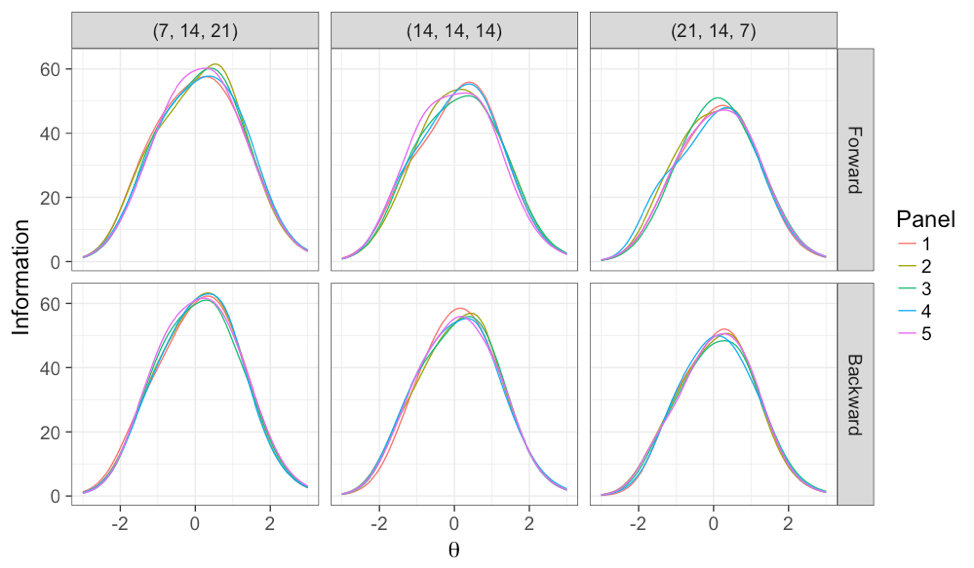
Response patterns were generated using the same R package. For each item, a random variable from Binomial(1, P(θ)) was simulated, where P(θ) is defined in Equation 1. If it was equal to 1, the simulee was said to answer the item correctly; otherwise the response was set to 0.

**Routing errors.** A routing error was said to occur when a simulee was routed to a module

***Figure 4.* Information functions of each panel for all 1-3-3 and 1-3-4 MSTs**

1. **1-3-3 MSTs**



**1-3-4 MSTs**

that did not provide the maximum information at his/her true *θ*. Routing errors resulting from transitions between stages 1 and 2 as well as stages 2 and 3 were both analyzed. Thus, a simulee was classified as misrouted at any routing point if it was assigned to a different module based on the *θ* estimate than it would have been assigned based on the true *θ* that generated the response pattern.

## Computerized Adaptive Test

The CAT administration was simulated using the *catR* package in R (Magis & Barrada, 2017). To ensure a fair comparison, each MST condition was compared with a CAT using an item bank that was comprised of the items selected for that particular MST condition. Items for all five panels were included, resulting in CAT item banks of from 420 to 665 items. Figure 5 shows the information and SEM functions for the CAT item banks based on the items selected for the MSTs in each condition. Identical to the MST conditions, the CATs had a fixed test length of 42 items and used MLE for *θ* estimation with a range of (−3.5, 3.5). The initial *θ* estimates used in selecting the first item were set by a random number generated from Unif(−0.5, 0.5). Items were selected based on the maximum Fisher information criterion.

**Evaluation Criteria**

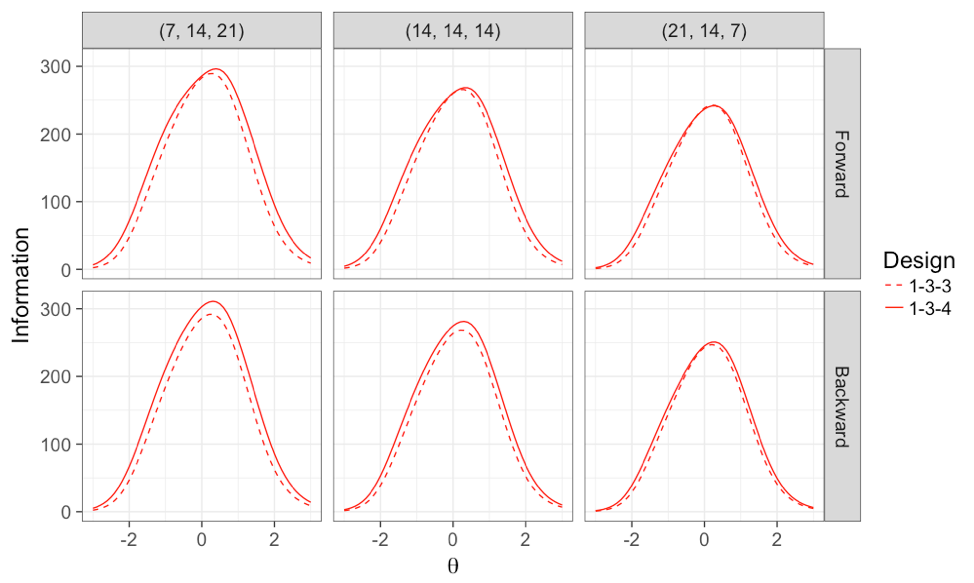
The measurement precision and efficiency of CATs and MSTs were compared across all manipulated conditions and separately for each number of MST routing errors, so that the effect of routing errors on measurement performance could be evaluated.

Precision. Following the practice of similar studies (Chen, 2010; Keng, 2008; Wang, 2017), the mean bias and root mean squared error (RMSE) were calculated to evaluate the recovery of true *θ*s. These two statistics were defined as

where is the true *θ* for simulee *j*, is the final *θ* estimate for simulee *j*, and *N* is the number of simulees. Each test design was also assessed in terms of the standard error of measurement (SEM)

***Figure 5.* Information and standard error of measurement (SEM)  
functions of CAT item banks**

1. **Information**



1. **SEM**

of the final *θ* estimate. The SEM for simulee *j* was obtained by

where is defined in Equation 1 and *n* is the number of items administered to examinee *j*.

Efficiency. The efficiency of each MST test design was evaluated in terms of how much reduction in test length could be achieved by “re-administering” the items as its corresponding CAT using the full item bank available to each MST: for each simulee, the SEM of the final *θ* estimate in an MST was used as a baseline; the number of items was determined that the simulee needed to reach the same SEM in a CAT of the same condition. In order to assess the differences across the *θ* scale, all evaluative statistics were calculated conditional on *θ*.

**Results**

**Routing errors**

Table 3presentsthe percentage of routing errors for each MST design. Percentage of routing errors ranged from a minimum of 8.4% for a forward assembled MST with half the items (21) in the first stage to 24.3% for backward assembled MSTs with 7 items in the first stage. The average number of routing errors was 16.3%. Overall, 1-3-3 MSTs had a lower average percentage of routing errors (11.5% and 14.0%) than 1-3-4 MSTs (18.5% and 21.1%), likely because there were more branches in the third stage of 1-3-4 MSTs. For 1-3-3 MSTs, allocating most items in the first stage also resulted in a lower percentage of routing errors (8.4% and 10.2%) than allocating the same number of items in all stages (11.0%) and allocating most items in the last stage (13.6%). For assembly priority, forward assembled MSTs had a lower percentage (15.0%) than backward assembled MSTs (17.6%), which was expected as they had items of higher quality at the routing stage. The largest percentage of routing errors (15% to 24.3%) was observed for the (7, 14, 21) design under both forward and backward assembly.

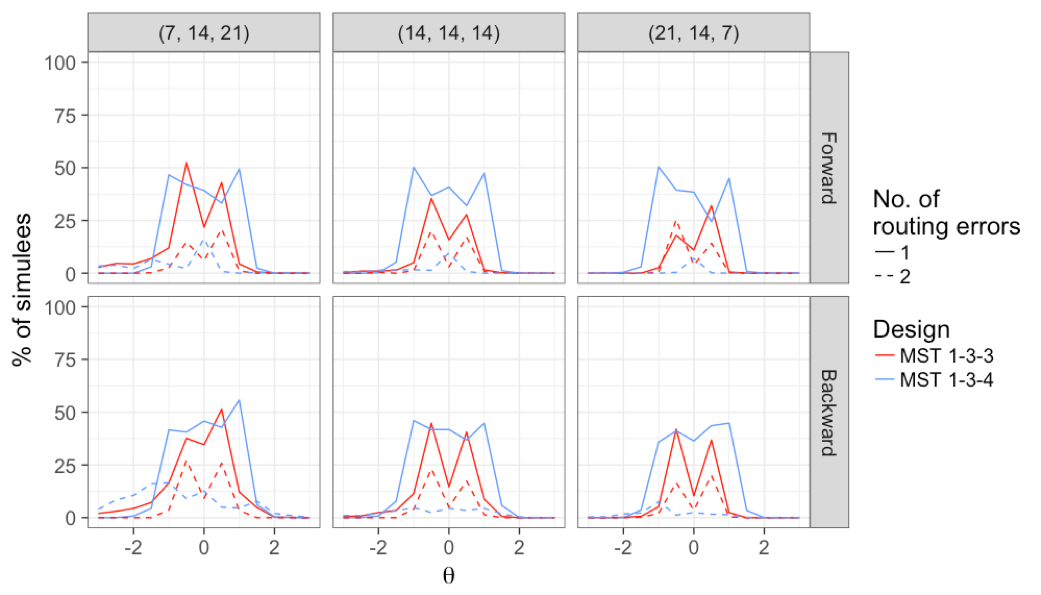
Figure 6 shows the percentage of routing errors conditional on *θ*. Most routing errors occurred in the center of the *θ* scale. When the simulee’s *θ* was at the *θ* anchor, the percentage of routing

**Table 3. Percentage of routing errors for each MST design**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Forward assembly | | Backward assembly | |  |
| Item allocation | 1-3-3 | 1-3-4 | 1-3-3 | 1-3-4 | Average |
| (7, 14, 21) | 15.0% | 20.9% | 18.1% | 24.3% | 19.6% |
| (14, 14, 14) | 11.0% | 18.0% | 13.6% | 20.1% | 15.7% |
| (21, 14, 7) | 8.4% | 16.7% | 10.2% | 19.0% | 13.6% |
| Average | 11.5% | 18.5% | 14.0% | 21.1% | 16.3% |

errors was relatively low. But when the simulee’s *θ* was in between two *θ* anchors, the percentage of routing errors elevated. Comparing the 1-3-3 and 1-3-4 designs, the 1-3-4 designs misrouted simulees over a wider range of *θ*. It is notable that for almost all MST designs, there were instances of approximately 50% of simulees with one routing error at some *θ* levels, particularly for the 1-3-4 design. For the backward assembly approach, the number of routing errors decreased as the number of items in the routing test increased.

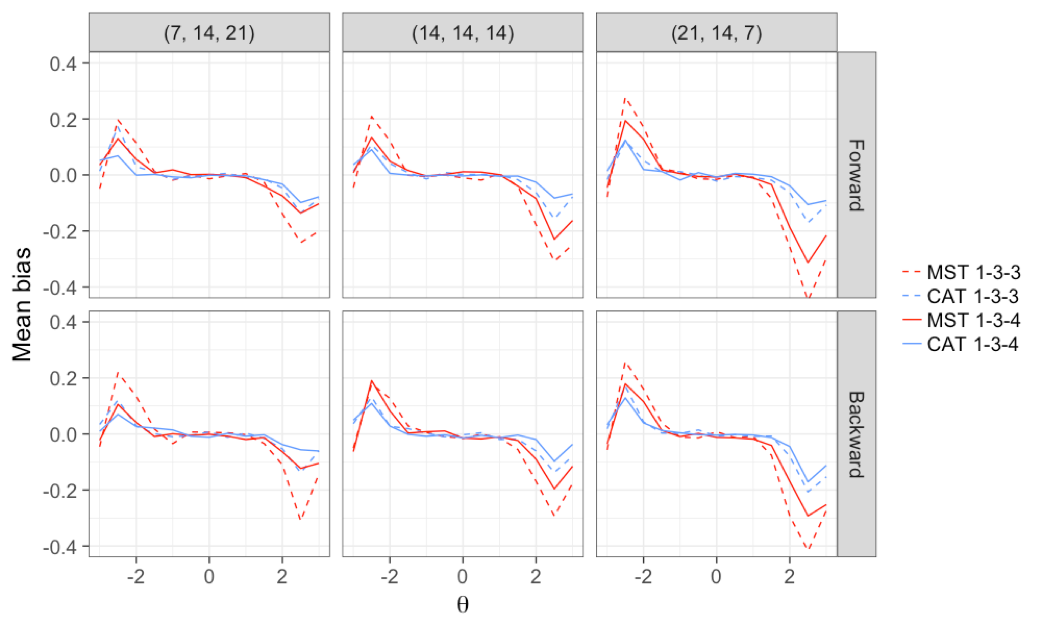
**Figure 6. Mean conditional percentage of simulees with routing errors**



**Mean Bias**

As illustrated in Figure 7, MSTs and CATs of all conditions performed equally well near the center of the *θ* scale, as their mean biases were all close to 0. However, the MSTs tended to

**Figure 7. Mean conditional bias for all MST conditions**

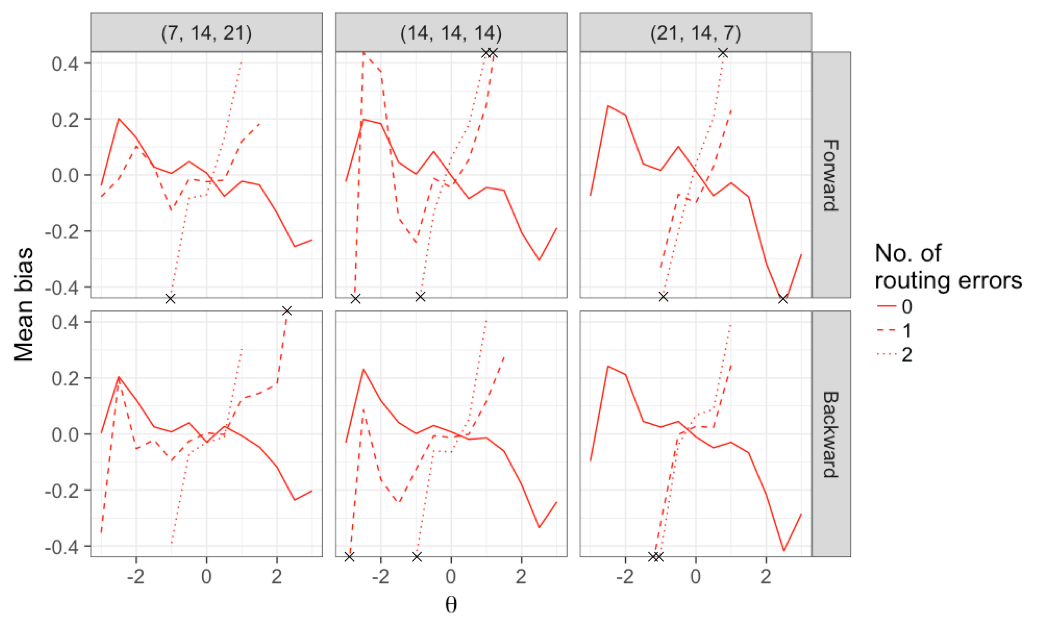


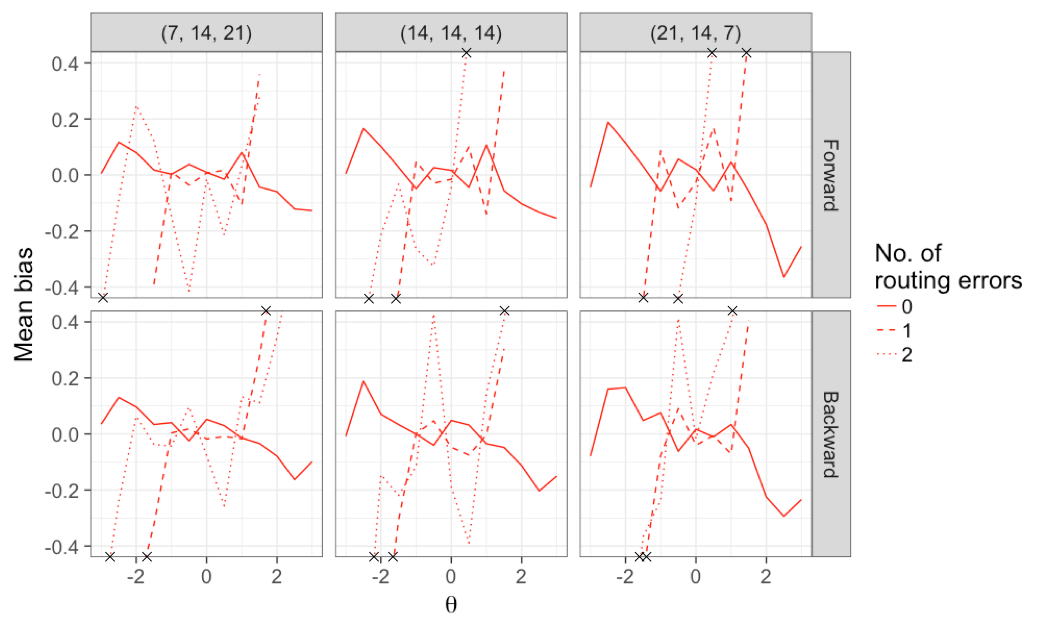
underestimate more than the CATs at low *θ* levels and overestimate at high *θ* levels. The underestimation and overestimation were larger for 1-3-3 MSTs than 1-3-4 MSTs, also for MSTs that had more items in the earlier stages than MSTs that had more items in the later stages. There appeared to be no effect of assembly priority.

The results of conditional mean bias grouped by the number of routing errors are shown in Figure 8. The results show considerable bias in the s for MSTs across a relatively wide range of *θ* (an x in these figures indicates a value beyond the plotted range of the y axis). Interestingly, the underestimation at low *θ*s and overestimation at high *θ*s occurred only for simulees with no misrouting. Simulees with one or two misroutings displayed the opposite trend. This explains the zero bias near the center of the *θ* scale in Figure 7—the positive and negative bias cancelled out when the biases of simulees with different numbers of misroutings were averaged. For the 1-3-3 design (Figure 8a), bias for simulees with two misroutings was generally higher than for those with a single misrouting, but was concentrated more toward the center of the *θ* scale. For the 1-3-4 design (Figure 8b), the same pattern was evident for the (21, 14, 7) MST but a reverse pattern

**Figure 8. Mean conditional bias grouped by the number of routing errors   
for 1-3-3 and 1-3-4 MSTs**

1. **1-3-3 MSTs**

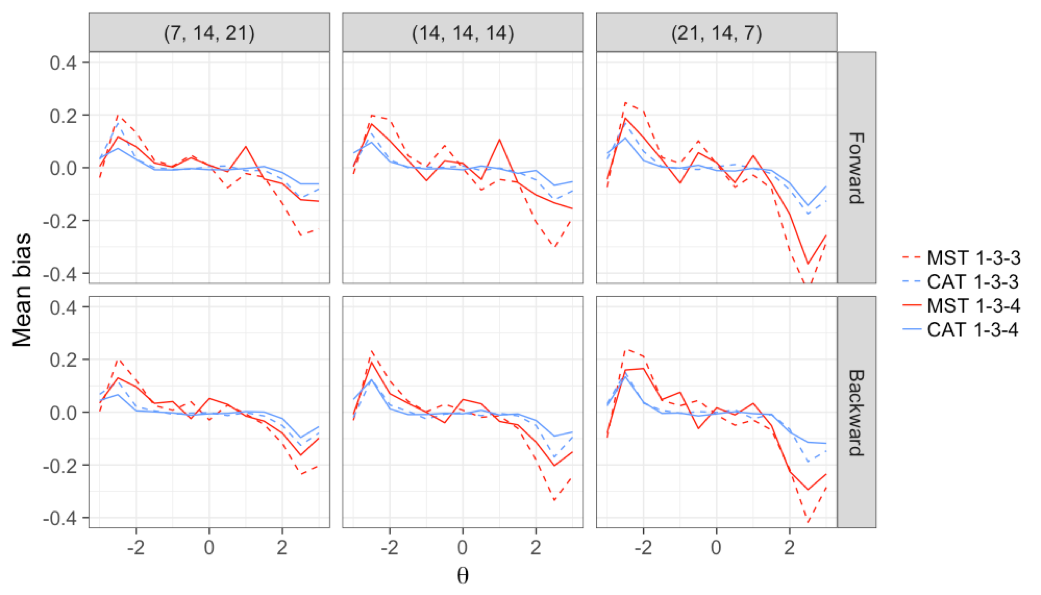


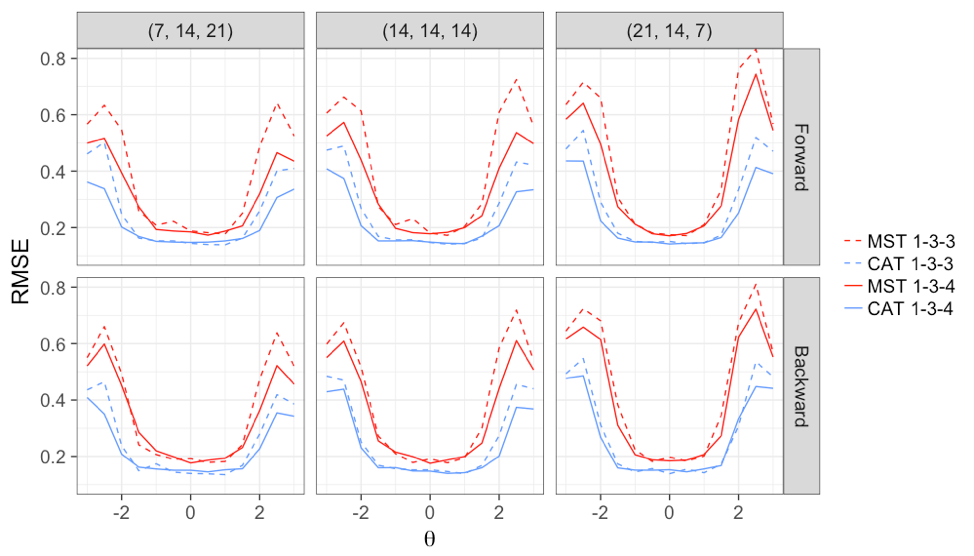
1. **1-3-4 MSTs**

emerged for MSTs with fewer items in the first stage.

To further evaluate the effect of routing errors, the results of conditional mean bias for MSTs with zero routing errors and all CATs are shown in Figure 9. After removing the routing error effect, the mean biases of MSTs were not as small as shown in Figure 7, particularly near the center of

**Figure 9. Mean conditional bias for MSTs with zero routing errors and CATs**



**Figure 10. Mean conditional RMSE for all MST conditions**

the *θ* scale, with the most bias evident for the 1-3-3 MST.

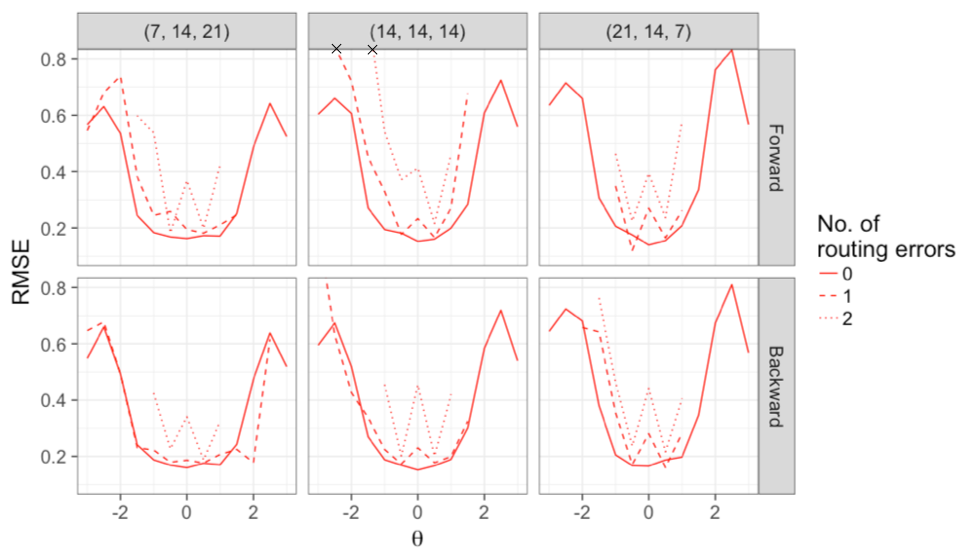
**RMSE**

Figure 10 shows the conditional RMSE for all conditions. Both CATs and MSTs performed the best near the center of the *θ* scale and poorest at the extremes, reflecting the information structure of the item banks (Figure 4). All CATs demonstrated a lower RMSE than their corresponding MSTs throughout the entire *θ* scale. This disparity was especially large at the extremes. The differences between each MST design were not obvious at the center of the *θ* scale, but at the extremes, 1-3-4 MSTs resulted in smaller RMSEs than 1-3-3 MSTs. A smaller number of items in the first stage (i.e., 7, 14, 21 versus 21, 14, 7) led to smaller RMSEs at the extremes. There was no obvious effect of assembly priority. As shown in Figure 11, having more routing errors tended to produce higher RMSEs, especially near the center of the *θ* scale for simulees with two misroutings. Figure 12 presents the conditional RMSE for MSTs with no routing errors and CATs. CATs still performed better at extreme *θ*s after the effect of misrouting was removed in the MSTs.

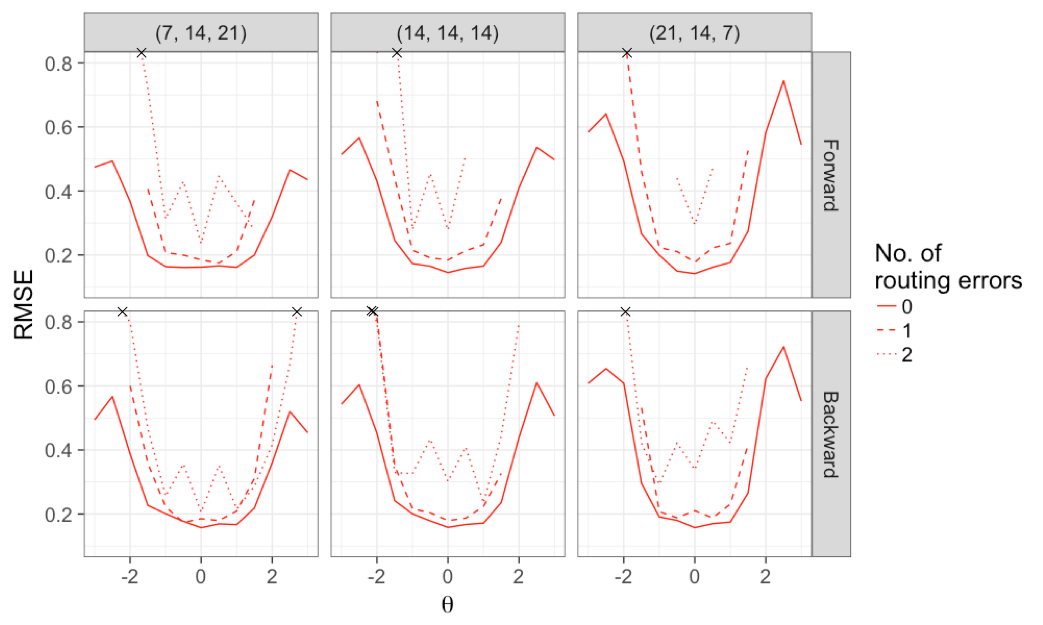
**Mean SEM**

The conditional mean SEM for all conditions is displayed in Figure 13. The CATs again performed better than their corresponding MSTs across the entire *θ* scale. Although the deficiency of the MSTs was small near the center of the scale, the MSTs exhibited a larger SEM toward the extreme ends of the scale, especially at the lower end. Differences among the MSTs were observed only in the extreme *θ*s. The 1-3-4 designs tended to have smaller SEMs than the 1-3-3 designs. Having more items allocated to later stages (7, 14, 21) resulted in smaller SEMs. Backward assembled MSTs had slightly smaller SEMs than forward assembled MSTs. As illustrated in Figure 14, having more misroutings did not have much effect on the SEMs for the 1-3-3 design (Figure 14a) but had a larger effect for the 1-3-4 design (Figure 14b), particularly for simulees that had two routing errors. Hence, the differences between Figures 13 and 15, which shows conditional SEMs for those simulees with no routing errors, were minimal, but MSTs without routing errors still had larger RMSEs throughout the *θ* range than did CATs using the same item banks.

**Figure 11. Mean conditional RMSE grouped by the number of routing errors   
for 1-3-3 and 1-3-4 MSTs**

**a. 1-3-3 MSTs**

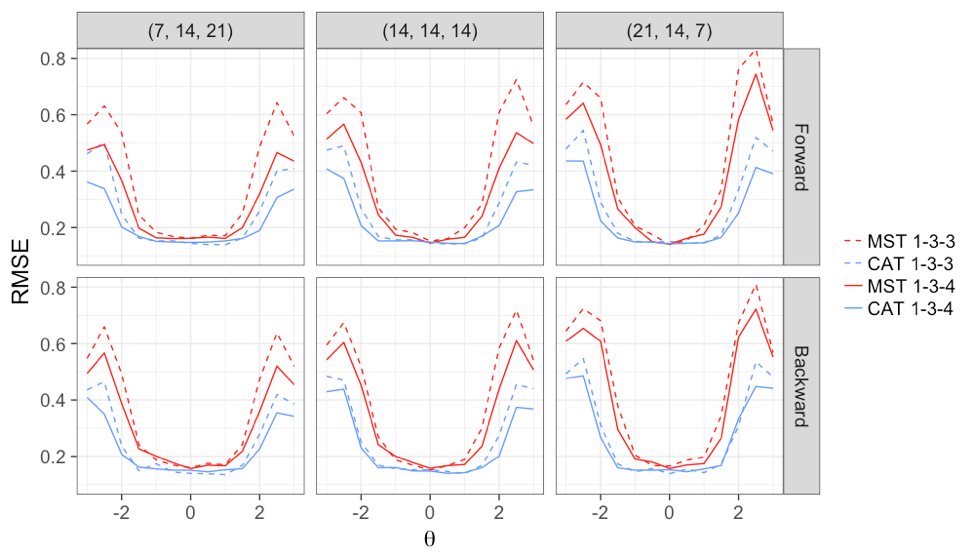
**b. 1-3-4 MSTs**



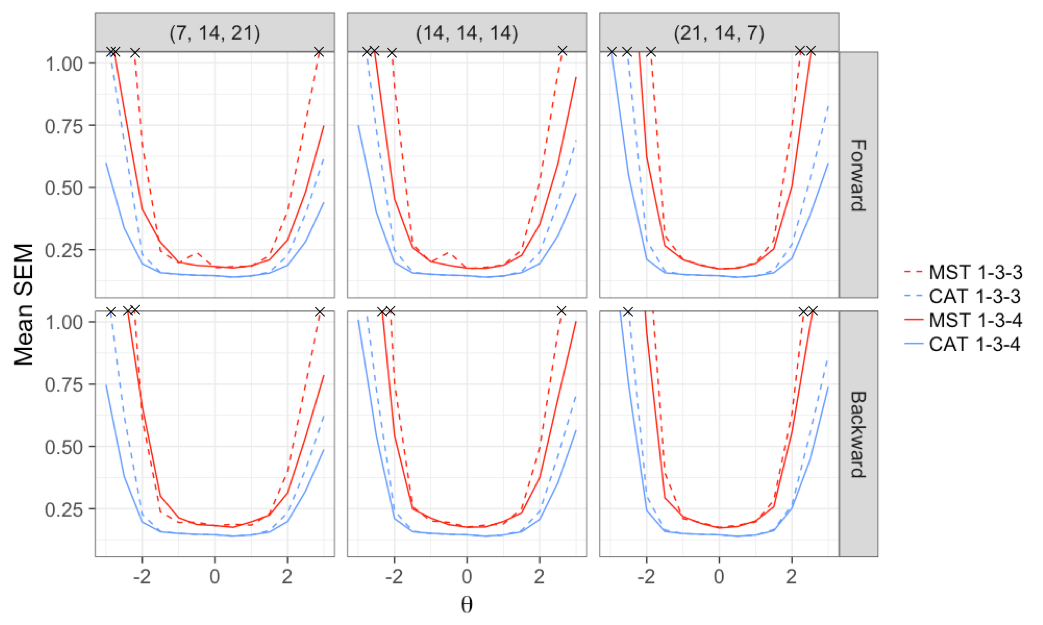
**Test length**

Figure 16 provides the conditional plot for how long CATs needed to be in order to reach the same SEM as their corresponding MSTs. Table 4 shows these results combined across θ levels. The CATs were more efficient than the MSTs overall—on average they needed about half the

**Figure 12. Mean conditional RMSE for MSTs with zero routing errors and CATs**



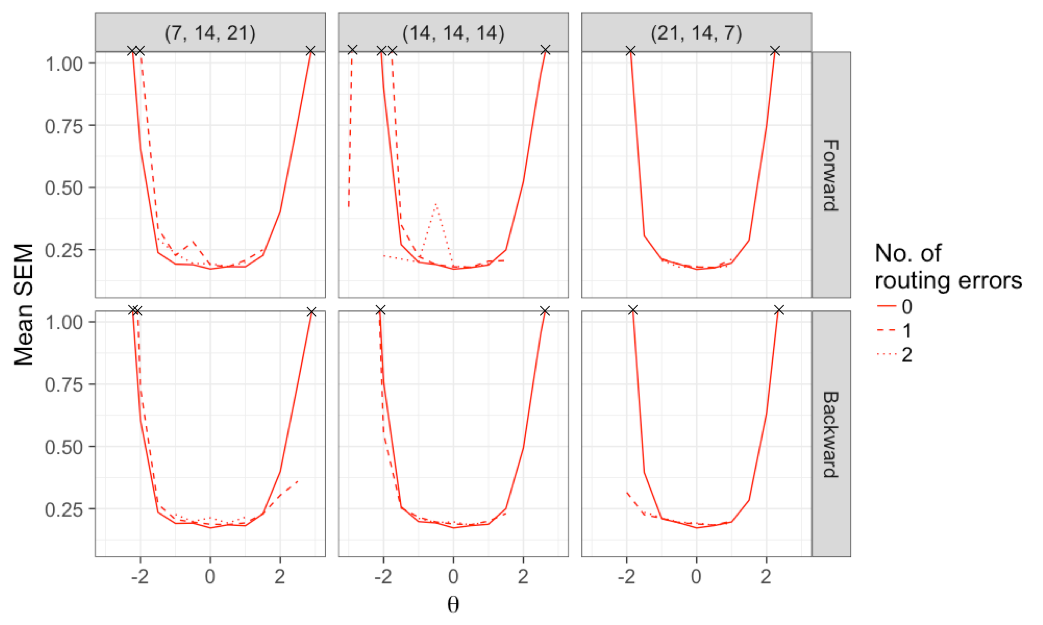
**Figure 13. Mean conditional SEM for all conditions**

****

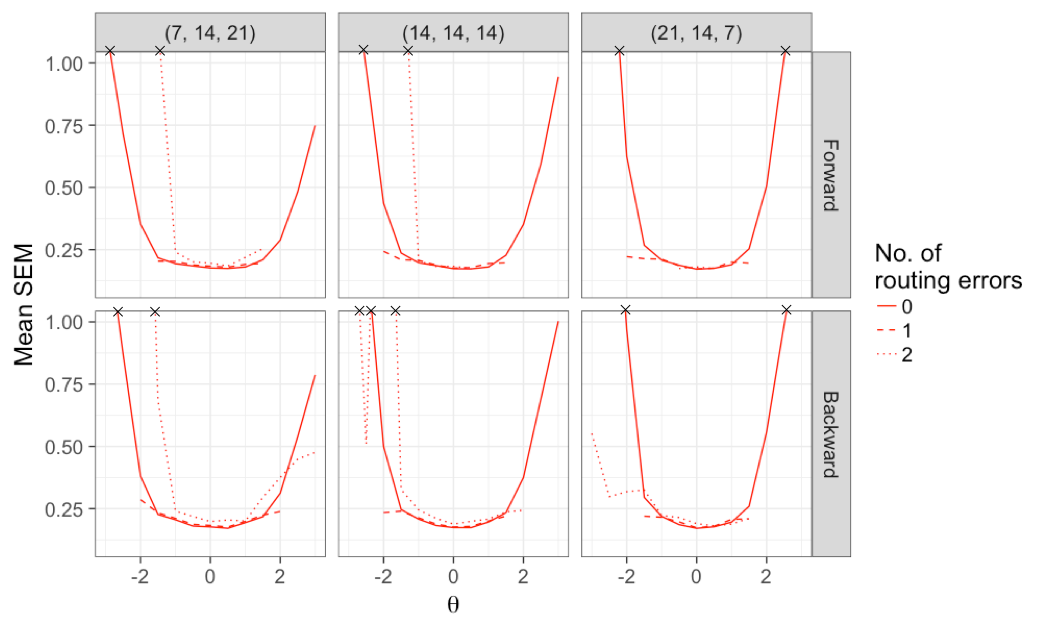
number of items (20.4 items) to achieve the same SEM as their corresponding 42-item MSTs across all θ levels. The number of items needed was smaller for extreme θs than θs in the middle of the scale. For θ around 0, CATs required an average of 27 to 30 items to match the average SEMs of the 42- item MSTs, with a slightly larger number of items for the MSTs with 21 items in the first

***Figure 14.* Mean conditional SEM grouped by the number of routing errors   
for 1-3-3 and 1-3-4 MSTs**

1. **1-3-3 MSTs**

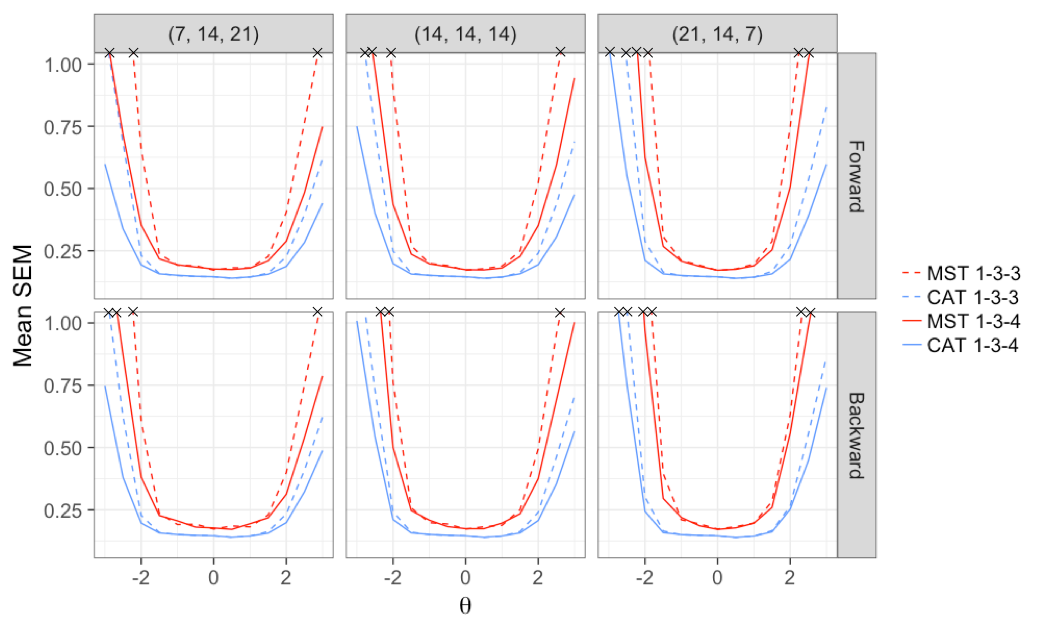
****

1. **1-3-4 MSTs**

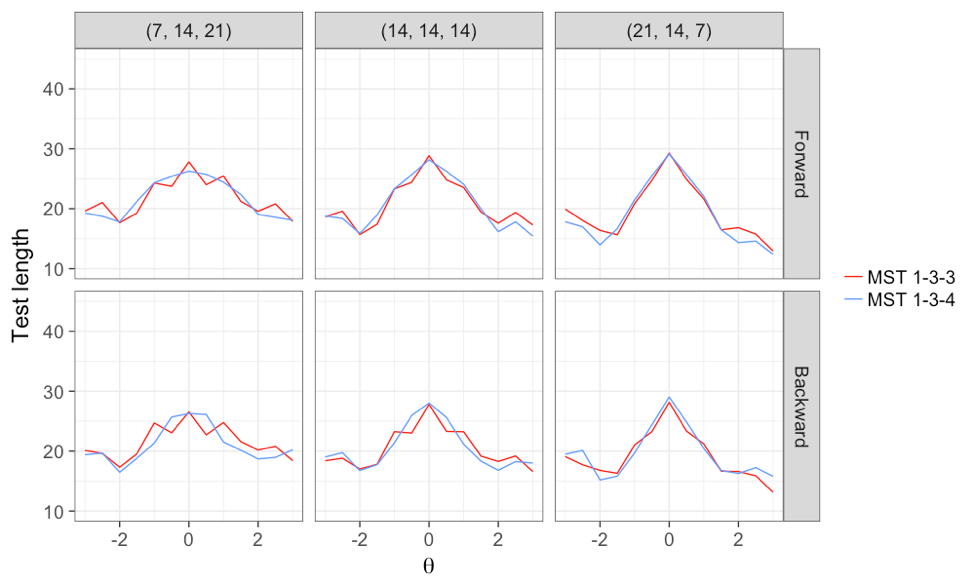


stage. Differences among MST designs were not obvious. As shown in Figure 17, there was a small trend that more routing errors in MSTs required fewer items in CATs, in the middle of the θ scale. Figure 16 shows the effect of number of routing errors on matching tests lengths; effects were particularly notable for 1-3-4 MSTs. Finally, Table 4 shows that neither number of routing errors

**Figure 15. Mean conditional SEM for MSTs with zero routing errors and CATs**



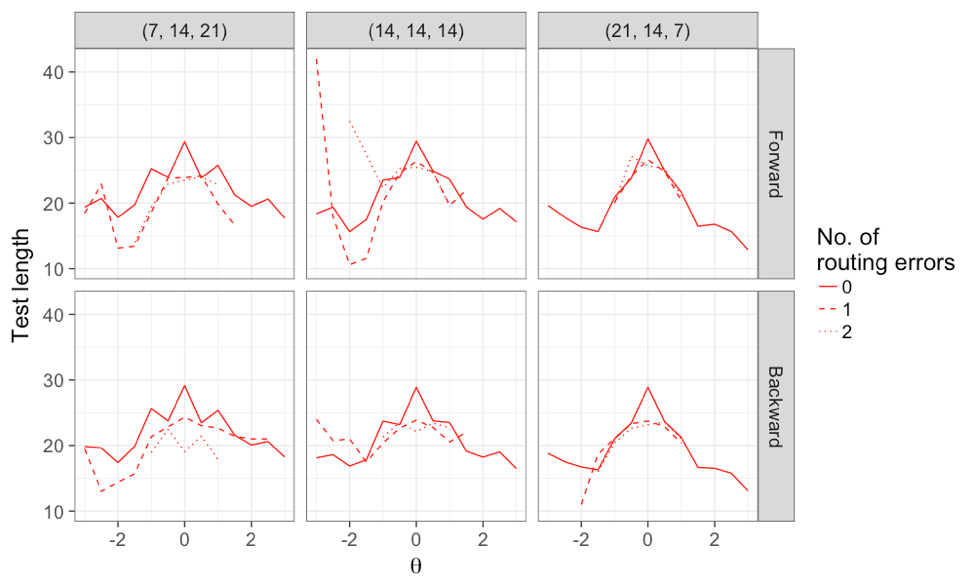
**Figure 16. Mean conditional test length required by CATs to achieve the same SEM   
as their corresponding MSTs for all conditions**



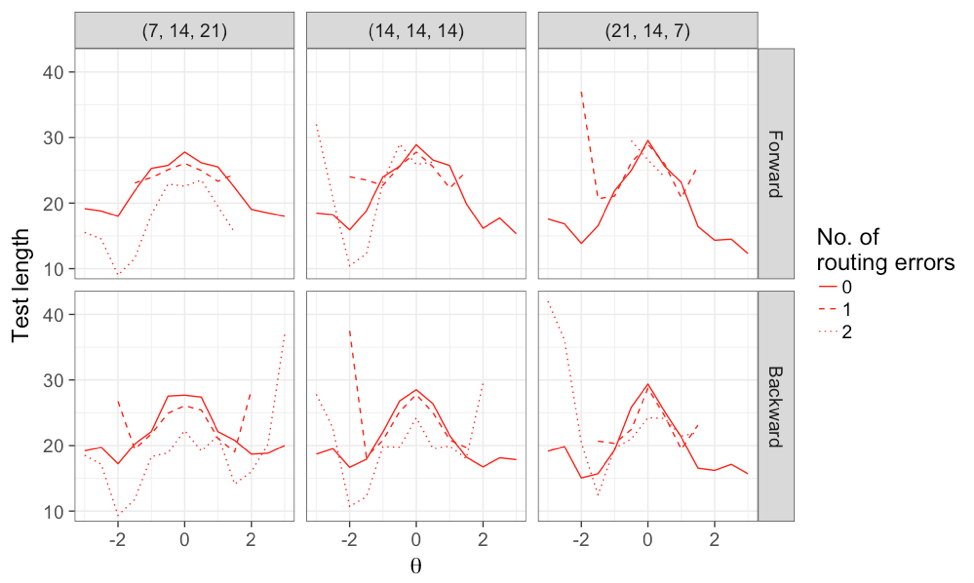
nor MST design had any major overall effect on the relationship between CAT test length and SEMs. On average, CATs required about 20 items to obtain SEMs comparable to those of the 42-item MSTs.

***Figure 17.* Mean conditional test length required by CATs to achieve the same SEM as their corresponding MSTs, grouped by the number of routing errors for 1-3-3 and 1-3-4 MSTs**

1. **1-3-3 MSTs**



1. **1-3-4 MSTs**



**Table 4. Mean test length required by CATs to achieve   
the same SEM as their corresponding MSTs for all conditions**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  | Number of   routing errors | | |  |
| Test structure | Assembly priority | Item allocation | 0 | 1 | 2 | Average |
| 1-3-3 | Forward | (7, 14, 21) | 21.7 | 21.6 | 21.3 | 21.7 |
|  |  | (14, 14, 14) | 20.7 | 20.4 | 21.4 | 20.7 |
|  |  | (21, 14, 7) | 19.0 | 19.9 | 19.8 | 19.4 |
|  | Backward | (7, 14, 21) | 21.4 | 21.9 | 21.0 | 21.4 |
|  |  | (14, 14, 14) | 20.5 | 20.0 | 19.7 | 20.4 |
|  |  | (21, 14, 7) | 19.1 | 19.0 | 19.2 | 19.1 |
| 1-3-4 | Forward | (7, 14, 21) | 21.7 | 21.1 | 21.9 | 21.6 |
|  |  | (14, 14, 14) | 20.7 | 20.5 | 20.9 | 20.6 |
|  |  | (21, 14, 7) | 18.9 | 19.1 | 19.9 | 19.0 |
|  | Backward | (7, 14, 21) | 21.0 | 20.7 | 21.3 | 21.0 |
|  |  | (14, 14, 14) | 20.5 | 20.2 | 21.0 | 20.5 |
|  |  | (21, 14, 7) | 19.5 | 19.6 | 19.7 | 19.6 |
| Average |  |  | 20.4 | 20.4 | 20.7 | 20.4 |

**Discussion and Conclusions**

**Misrouting**

The results of the current study demonstrate that, when averaging across *θ*s and all investigated conditions, routing errors occurred 16.3% of the times, with a range of 13.6% to 19.6%, depending on item allocation in the MSTs. Placing more items in the first stage of an MST resulted in the fewest routing errors, whereas placing more items in the third stage test resulted in the most routing errors. The percentage of routing errors generally supported the earliest two-stage research that estimated about 20% routing errors (Angoff & Huddleston, 1958). The current study appears to be the first that examined the effect of *θ* level on routing errors. As expected, most routing errors were observed when the simulee’s true *θ* was close to a *θ* anchor: under these circumstances routing errors occurred for about 50% of examinees, a value that exceeded the 40% misrouting rate reported by Cleary et al. (1969).

Routing errors were also found to adversely affect bias, RMSE and, to some extent, SEM. When MSTs were compared to CATs, based on data that included both simulees with no misroutings and those who were misrouted, there appeared to be very little effect of routing errors on these criteria. But when those who were misrouted were separated from those who were not, the picture changed dramatically, especially for bias. Misrouting resulted in extreme directional bias at different *θ* levels, which were cancelled when the data were aggregated across misrouting conditions. Misrouted simulees with *θ*s near the anchor points had very high levels of bias whereas those who were not misrouted displayed bias in the opposite direction. Although conditional bias functions without regard to misrouting differed only in the extremes for MSTs versus CATs, a comparison of those only with no misroutings in their MSTs versus CATs showed results less favorable to MSTs, with a more erratic pattern of conditional bias and RMSE—eliminating simulees with misroutings increased the difference in conditional bias and RMSE functions between MSTs and CATs.

SEM showed some effect due to misrouting, particularly when the results were examined taking into account routing errors. However, there was little difference in conditional SEM functions for the group without misroutings as compared to the total group that included misrouting. This result contrasts sharply with those obtained for bias and RMSE. This suggests that the SEM of the *θ* estimates as computed from Equation 6 is not an adequate criterion for evaluating the effects of misrouting, or by extension the comparable performance of different methods of testing. This is because the SEM is computed as a confidence interval around an estimate of *θ*, and it implicitly assumes that the estimate is a reasonably accurate approximation to the true *θ*—that is, that *θ* is unbiased. To the extent that *θ* is biased, as was demonstrated in this study, the *θ* estimate is not an adequate representation of true *θ* and, hence, the confidence interval represented by the SEM is essentially invalid. As Lord (1983) noted, the bias in MLEs arises from low test information at the examinee’s *θ* estimate: when a test provides low information for an examinee, the result is a biased *θ* estimate (Weiss & von Minden, 2011). It is this low information that occurs in MSTs with each routing error that causes the elevated bias and RMSE in the *θ* estimates, and invalidates the SEM as an indicator of measurement precision.

**MST Design**

The current study also provides guidelines for designing an MST, where mixed results have previously been found. Much of the MST research (e.g. Wang, 2017; Zheng & Chang, 2015) has focused on the 1-3-3 structure. This design was, however, generally shown to be inferior to the 1-3-4 design in the current study, possibly because it did not provide the potential to improve the *θ* estimates obtained at the second stage for examinees with extreme abilities. This corroborates Patsula’s (1999) study which showed that increasing the number of modules led to more accurate *θ* estimation. The current study demonstrated thatplacing more items in the last stage tended to produce more precise measurement, supporting the rationale that examinees should receive more items closely matched to their ability. However, this outcome is contrary to Patsula (1999) and Zheng et al. (2012) who found little effect for item allocation. As for assembly priority, no obvious impact was found in the current study, which was consistent with the findings by Wang (2017) but differed from Zheng et al. (2012) who found that backward assembled MSTs outperformed forward assembled MSTs in terms of classification accuracy. A possible explanation for such discrepancy is the difference in evaluation metrics (classification vs. measurement accuracy).

However, a slightly different picture emerged when taking routing errors into account. Under both forward and backward assembly, 1-3-3 designs resulted in up to 50% fewer misroutings than for 1-3-4 designs, likely due to fewer routing options. The smallest percentage of misroutings occurred for a forward 1-3-3 MST with 21 items in Stage 1 and the largest (by almost three times) for a backward 1-3-4 design with seven items in the first stage. Thus, the choice of design for an MST should focus on minimizing the number and effects of routing errors, to provide maximally effective measurements.

**CAT Versus MST**

Previous research found that CATs tend to outperform MSTs (e.g., Kim & Plake, 1993; Wang, 2017; Yan et al., 2014). The current study improved on prior research by using the same item bank and a more appropriate three -parameter IRT model. In general, when the misrouting effect was included in the analyses, the measurement performance for MSTs, including bias, RMSE and SEM, approximated that of CATs using the same item bank only in the middle *θ* range. Outside that narrow *θ* range, CATs yielded superior performance than MSTs. In terms of efficiency, CATs were able to achieve the same SEMs as MSTs with an average 50% fewer items. These results remained true even when the misrouting effect in MSTs was excluded.

**Limitations and Future Research Directions**

A possible limitation to this study was that no content balancing or item exposure control was implemented. In practice, these constraints are often imposed to ensure content coverage and test security. Future research can assess the impact of these practical constraints on the functioning of MSTs and CATs. In addition, the *θ* anchors in the last two stages of the 1-3-3 design used in this study were limited to (1, 0, 1). Future studies can examine whether using a wider set of *θ* anchors e.g. (2, 0, 2) in the last stage can overcome the poor estimation in the extreme *θ* ranges and/or reduce the number of routing errors or their effects on measurement precision*.*

**Conclusions**

Routing errors in MSTs were found to adversely affect measurement quality. Results of routing error analyses imply a need to seek ways to minimize the frequency or effects of routing errors. More research is necessary to evaluate solutions to the misrouting problem, since the present results indicate that the (7,14,21) design had the smallest percent of routing errors whereas the (21,14,7) design had the best performance in terms of lowest—but substantial—conditional bias and RMSE. MSTs have been proposed as a means of controlling content balance and other test assembly characteristics (Yan et al., 2014), but the present results raise the issue of whether those objectives are worth the tradeoffs in measurement accuracy, precision, and efficiency attainable by CATs. MSTs might also not be the best choice for tests intending to measure individual differences rather than making pass-fail decisions, because the *θ* estimates of examinees with high or low *θ*s will be both biased and imprecise, as well as those of examinees who have been misrouted.

**References**

[Angoff, W. H, & Huddleston, E. M](http://iacat.org/biblio?f%5Bsearch%5D=angoff&f%5Bauthor%5D=1415). (1958). [*The multi-level experiment: A study of a two-level test system for the College Board Scholastic Aptitude Test*](http://iacat.org/content/multi-level-experiment-study-two-level-test-system-college-board-scholastic-aptitude-test). Princeton NJ: Educational Testing Service.

[Betz, N. E.](http://iacat.org/biblio?f%5Bsearch%5D=betz&f%5Bauthor%5D=755&s=author&o=asc), & [Weiss, D. J.](http://iacat.org/biblio?f%5Bsearch%5D=betz&f%5Bauthor%5D=1733&s=author&o=asc). (1973). [*An empirical study of computer-administered two-stage ability testing* (Research Report 73-4)](http://iacat.org/content/empirical-study-computer-administered-two-stage-ability-testing-research-report-73-4). Minneapolis: Department of Psychology, Psychometric Methods Program.

[Betz, N. E.](http://iacat.org/biblio?f%5Bsearch%5D=betz&f%5Bauthor%5D=755&s=author&o=asc), & [Weiss, D. J.](http://iacat.org/biblio?f%5Bsearch%5D=betz&f%5Bauthor%5D=1733&s=author&o=asc). (1974). [*Simulation studies of two-stage ability testing* (Research Report 74-4)](http://iacat.org/content/simulation-studies-two-stage-ability-testing-research-report-74-4). Minneapolis: Department of Psychology, Psychometric Methods Program.

Chen, L. Y. (2010). *An investigation of the optimal test design for multi-stage test using the generalized partial credit model* (Unpublished doctoral dissertation). The University of Texas at Austin.

[Cleary, T. A.](http://iacat.org/biblio?f%5Bsearch%5D=cleary&f%5Bauthor%5D=505), [Linn, R. L.](http://iacat.org/biblio?f%5Bsearch%5D=cleary&f%5Bauthor%5D=321), & [Rock, D. A.](http://iacat.org/biblio?f%5Bsearch%5D=cleary&f%5Bauthor%5D=504). (1969). [An exploratory study of programmed tests](http://iacat.org/content/exploratory-study-programmed-tests). *Educational and Psychological Measurement*, *28*, 345-360.

Davis, L. L., & Dodd, B. G. (2003). Item exposure constraints for testlets in the verbal reasoning section of the MCAT. *Applied Psychological Measurement, 27*(5), 335-356.

Hambleton, R. K., & Xing, D. (2006). Optimal and nonoptimal computer-based test designs for making pass–fail decisions. *Applied Measurement in Education, 19*(3), 221-239.

[Hornke, L. F](http://iacat.org/biblio?f%5Bsearch%5D=hornke&f%5Bauthor%5D=397). (1979). [Four realizations of pyramidal adaptive testing](http://iacat.org/content/four-realizations-pyramidal-adaptive-testing). *Programmed Learning and Educational Technology*, *16*, 164-169.

Jodoin, M. G., Zenisky, A., & Hambleton, R. K. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Applied Measurement in Education, 19*(3), 203-220.

Keng, L. (2008). *A comparison of the performance of testlet-based computer adaptive tests and multistage tests* (Unpublished doctoral dissertation). The University of Texas at Austin.

Kim, S., & Moses, T. (2014). An Investigation of the Impact of Misrouting Under Two-Stage Multistage Testing: A Simulation Study. *ETS Research Report Series, 2014*(1), 1-13.

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

[Larkin, K. C.](http://iacat.org/biblio?f%5Bsearch%5D=larkin&f%5Bauthor%5D=1480), & [Weiss, D. J.](http://iacat.org/biblio?f%5Bsearch%5D=larkin&f%5Bauthor%5D=1733). (1975). [*An empirical comparison of two-stage and pyramidal ability testing* (Research Report 75-1)](http://iacat.org/content/empirical-comparison-two-stage-and-pyramidal-ability-testing-research-report-75-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Linn, R. L., Rock, D. A. & Cleary, T. A. (1969). The development and evaluation of several programmed testing methods. *Educational and Psychological Measurement,* 1969, 129-146.

[Lord, F. M](http://iacat.org/biblio?f%5Bsearch%5D=lord&f%5Bauthor%5D=855). (1971a). [A theoretical study of the measurement effectiveness of flexilevel tests](http://iacat.org/content/theoretical-study-measurement-effectiveness-flexilevel-tests). *Educational and Psychological Measurement*, *31*, 805-813.

[Lord, F. M](http://iacat.org/biblio?f%5Bsearch%5D=lord&f%5Bauthor%5D=855). (1971b). [A theoretical study of two-stage testing](http://iacat.org/content/theoretical-study-two-stage-testing). *Psychometrika*, *36*, 227-242.

[Lord, F. M](http://iacat.org/biblio?f%5Bsearch%5D=lord&f%5Bauthor%5D=855). (1971c). [Robbins-Monro procedures for tailored testing](http://iacat.org/content/robbins-monro-procedures-tailored-testing). *Educational and Psychological Measurement*, *31*, 3-31.

[Lord, F. M](http://iacat.org/biblio?f%5Bsearch%5D=lord&f%5Bauthor%5D=855). (1974). [*Practical methods for redesigning a homogeneous test, also for designing a multilevel test* (RB-74-30)](http://iacat.org/content/practical-methods-redesigning-homogeneous-test-also-designing-multilevel-test-rb-74-30). Princeton NJ: Educational Testing Service.

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability, *Psychometrika, 48,* 233-245.

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35*(3), 229-249.

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

Luo, X., & Kim, D. (2018). A top-down approach to designing the computerized adaptive multistage test. *Journal of Educational Measurement, 55*(2), 243-263.

Macken-Ruiz, C. L. (2008). *A comparison of multi-stage and computerized adaptive tests based on the generalized partial credit model* (Unpublished doctoral dissertation). The University of Texas at Austin.

Magis, D., & Barrada, J. R. (2017). Computerized adaptive testing with R: Recent updates of the package catR. *Journal of Statistical Software, 76*(1), 1-19.

Magis, D., Yan, D., & von Davier, A. (2018). mstR: An R package to generate multistage testing designs. Available at <https://cran.r-project.org/web/packages/mstR/index.html>

Mead, A. D. (2006). An introduction to multistage testing. *Applied Measurement in Education, 19*(3), 185-187.

Oranje, A., Mazzeo. J., Xu, X., & Kulick, E. (2014). A multistage testing approach to group-score assessments. In D. Yan, A.A. von Davier, & C. Lewis (Eds.). *Computerized multistage testing: Theory and applications*. Boca Raton, FL: CRC Press.

Patsula, L N. (1999). A comparison of computerized adaptive testing and multi-stage testing. (Unpublished doctoral dissertation). University of Massachusetts Amherst.

Rotou, O., Patsula, L., Steffen, M., & Rizavi, S. (2007). Comparison of multistage tests with computerized adaptive and paper-and-pencil tests. *ETS Research Report Series, 2007*(1).

Schnipke, D. L., & Reese, L. M. (1999). *A Comparison [of] Testlet-Based Test Designs for Computerized Adaptive Testing.* Law School Admission Council Computerized Testing Report. LSAC Research Report Series.

[Vale, C. D.](http://iacat.org/biblio?f%5Bsearch%5D=vale&f%5Bauthor%5D=1049), & [Weiss, D. J.](http://iacat.org/biblio?f%5Bsearch%5D=vale&f%5Bauthor%5D=1733). (1975). [*A study of computer-administered stradaptive ability testing* (Research Report 75-4)](http://iacat.org/content/study-computer-administered-stradaptive-ability-testing-research-report-75-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Wang, K. (2017). *A Fair Comparison of the Performance of Computerized Adaptive Testing and Multistage Adaptive Testing* (Unpublished doctoral dissertation). Michigan State University.

Wang, X. (2013). *An investigation on computer-adaptive multistage testing panels for multidimensional assessment* (Unpublished doctoral dissertation). The University of North Carolina at Greensboro.

[Waters, B. K](http://iacat.org/biblio?f%5Bsearch%5D=waters&f%5Bauthor%5D=516). (1977). [An empirical investigation of the stratified adaptive computerized testing model](http://iacat.org/content/empirical-investigation-stratified-adaptive-computerized-testing-model). *Applied Psychological Measurement*, *1*(1), 141-152.

[Weiss, D. J](http://iacat.org/biblio?f%5Bsearch%5D=74-5&f%5Bauthor%5D=1733). (1974). [*Strategies of adaptive ability measurement* (Research Report 74-5)](http://iacat.org/content/strategies-adaptive-ability-measurement-research-report-74-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

[Weiss, D. J.](http://iacat.org/biblio?f%5Bsearch%5D=betz&f%5Bauthor%5D=1733&s=author&o=asc), & [Betz, N. E.](http://iacat.org/biblio?f%5Bsearch%5D=betz&f%5Bauthor%5D=755&s=author&o=asc). (1973). [*Ability measurement: Conventional or adaptive?* (Research Report 73-1)](http://iacat.org/content/ability-measurement-conventional-or-adaptive-research-report-73-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Weiss, D. J., & Von Minden, S. (2011). Measuring individual growth with conventional and adaptive tests. *Journal of Methods and Measurement in the Social Sciences, 2,* 80-101.

Xing, D., & Hambleton, R. K. (2004). Impact of test design, item quality, and item bank size on the psychometric properties of computer-based credentialing examinations. *Educational and Psychological Measurement, 64*(1), 5-21.

Yan, D., von Davier, A. A., & Lewis, C. (2014). *Computerized multistage testing: Theory and applications*. Boca Raton, FL: CRC Press.

Zheng, Y., & Chang, H. H. (2015). On-the-fly assembled multistage adaptive testing. *Applied Psychological Measurement, 39*(2), 104-118.

Zheng, Y., Nozawa, Y., Gao, X., & Chang, H. H. (2012). *Multistage adaptive testing for a large-scale classification test: The designs, automated heuristic assembly, and comparison with other testing modes* (ACT Research Reports 2012-6). Retrieved from http://media.act.org/documents/ACT\_RR2012-6.pdf