**Effect of Routing Errors on the Psychometric Properties of Multistage Tests  
David J. Weiss and King Yiu Suen  
University of Minnesota**

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; Betz & Weiss, 1974; Cleary, Linn, & Rock, 1969; Linn, Rock, & Cleary, 1969, Weiss & Betz, 1973). 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 routed to a module that is not well matched with his/her true trait (*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 40-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 MSTs with a routing test followed by three second-stage tests 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. They found that the score differences associated with different paths were negligible and the results from the two conditions were almost indistinguishable. The design of this study did not allow for evaluating the number or proportion of routing errors. Rather, the authors’ conclusion that in their study “… the impact of misrouting was minimal,” based on an analysis of number=correct MST scores converted to an IRT metric.

Luo and Kim (2018) reported the only other study to explicitly address misrouting between MST modules. 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). 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* regions were most affected by routing errors.

A study by Han (2020, and this volume chapter 7) examined suboptimal routing (i.e., routing errors evaluated at the MST panel level) in the context of intersectional routing. In this approach, prior information on an examinee from other sources (e.g., tests that an examinee had previously taken) is used to select one of two or more initial routing tests, and the score from the selected routing test is then used for subsequent routing within the MST. Han evaluated the procedure with short tests (seven items per stage) within simulation studies that varied the validity of the prior information and then evaluated the results in terms of the percentage of suboptimal paths through the MST, with routing errors within the MST responsible for suboptimal paths. Results showed that the percentage of suboptimal paths varied from about 35% to 15%, with the percentage decreasing as the validity of the prior information increased. However, even with perfect initial routing to the first MST stage, there were 5% to 15% suboptimal paths, which reflect routing errors within the MST independent of the validity of the prior information. A real data study reported suboptimal routing for from 15% to 20% of examinees using the intersectional routing procedure.

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.

## Purpose of the Study

The primary objective of this study was to examine, in simulation, how routing errors affect MST performance, as evaluated by the bias, root mean square error, and standard errors of the *θ* estimates. In addition, typical MST design factors were varied to determine how the measurement performance of MSTs of different types was affected in the presence of routing errors.

# 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 pool*. This study used Monte Carlo simulation methods. For each simulee in each condition, responses were simulated for all items in the operational pool, then MSTs were applied to estimate *θ*.

## 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

## 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, von Davier, & Lewis, 2014), the present study constructed five panels to make the simulation more realistic.

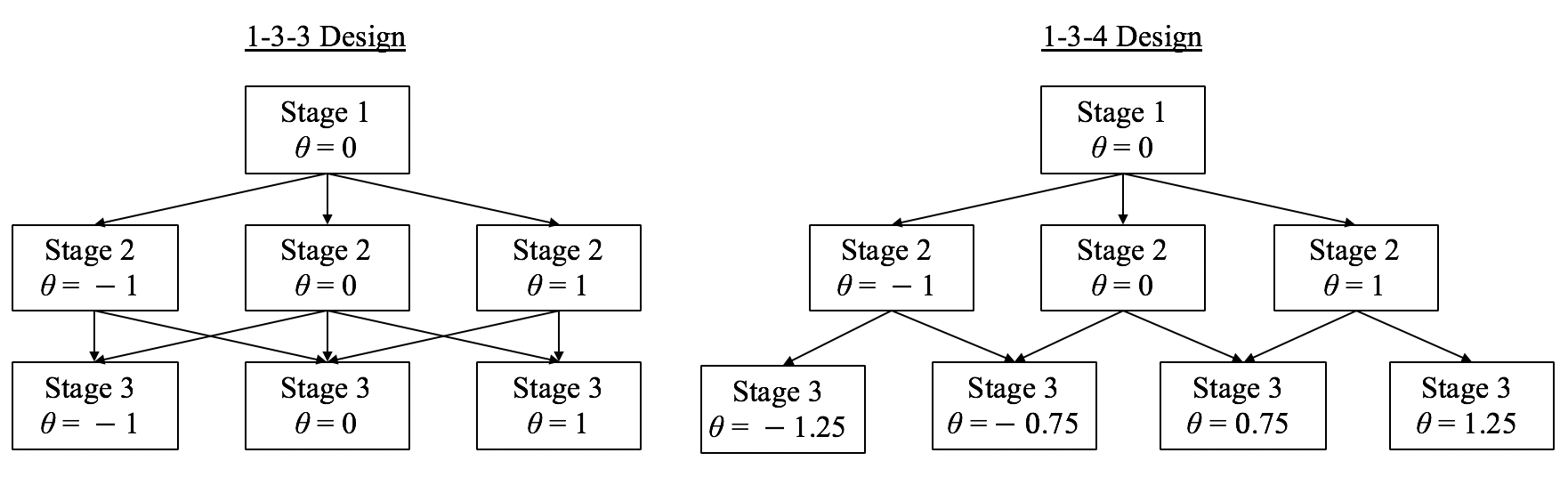
Conditions. Three MST design factors were manipulated to test which MST design yielded the best performance. This resulted in a total of 2 × 3 × 2 = 12 MST conditions, as shown in Table 1.

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

**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 |

which the module information was maximized are shown in Figure 1. 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).

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

### Item allocation. Three levels of item allocation were evaluated (Patsula, 1999). 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.

Assembly priority. The third factor was the assembly priority (forward and backward). In forward assembly, module assembly begins with Stage 1 and proceeds through the following stages. By contrast, backward assembly begins with the stage with most modules (typically the last stage) and proceeds assigning items through the earlier stages, with Stage 1 receiving its items last (Zheng, Nozawa, Gao and Chang, 2012).

### 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 item 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 recommended, 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.

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Parameter | 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) |

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 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 2 shows the module information functions of the 1-3-3 and 1-3-4 MSTs with equal numbers of items per stage, respectively, averaged over five panels.

**Figure 2. Module information functions for 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. Shape

   Description automatically generated**1-3-3 MST**
2. Diagram, shape

   Description automatically generated**1-3-4 MST**

Test administration. 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.

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 the random variable 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 occurred when a simulee was routed to a module that did not provide the maximum information at its 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.

**Evaluation Criteria**

The measurement precision of the 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. All evaluation criteria were computed conditional on *θ*. Mean bias and root mean squared error (RMSE) were calculated to evaluate the recovery of true *θ*s at each of the studied *θ* points. These two statistics were defined as

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

where is defined in Equation 2.

**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 routing targets in the third stage of 1-3-4 MSTs. For both 1-3-3 and 1-3-4 MSTs, allocating most items in the first stage also resulted in a lower percentage of routing errors (8.4% and 10.2% for 1-3-3 designs, and 16.7% and 19.0% for 1-3-4) than allocating the same number of items in all stages (11.0% and 13.6%, and 18.0% and 20,1%, repectively) and allocating most items in the last stage (15.0% and 18.1%, and 20.9% and 24.3%, repectively). For assembly priority, forward assembled MSTs had a lower percentage of routing errors ( average 15.0%) than backward assembled MSTs (17.6%), which was expected as they had items of higher quality at the routing stages. The largest percentage of routing errors (15% to 24.3%) was observed for the (7, 14, 21) design under both forward and backward assembly.

**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% |

Figure 3 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 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 (Figure 3b) misrouted simulees over a wider range of *θ*. However, the number of routing errors decreased with increasing numbers of items in the routing test for the 1-3-3 design (Figure 3a) but not for the 1-3-4 design (Figure 3b). 1-3-3 MSTs also had more simulees who had two routing errors than did the 1-3-4 MSTs. 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.

**Figure 3. Mean conditional percentage of simulees with routing errors**

A picture containing diagram

Description automatically generatedCalendar

Description automatically generated with low confidence **a. 1-3-3- MSTs b. 1-3-4 MSTs**

**Mean Bias**

Figure 4a shows mean bias across *θ* for each MST design (numerical values for all dependent variables are in Appendix Table A) averaged across routing error conditions. MSTs of all conditions performed equally well near the center of the *θ* scale for *θ* between -1.5 and 1.5, as their mean biases were all close to 0, but varied across *θ* levels. However, the MSTs tended to underestimate at low *θ* levels and overestimate at *θ* levels above the center of the θ scale—negative bias of about 0.10 and greater was observed for θ < -2.0 with values as high as -0.25 for a forward assembled 1-3-3 (21, 14, 7) design at θ = -2.5. On the positive end of θ, higher levels of bias occurred at θ = 2.0 and above, with the values of bias in that region generally higher than those observed for negative θs. The largest value of bias observed was bias = 0.47 at θ = 2.5 for the same forward assembled 1-3-3 (21, 14, 7) design. 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 was no general effect of assembly priority.

The results of conditional mean bias for MSTs with zero routing errors are shown in Figure 4b for 1-3-3 and 1-3-4 MSTs. Comparison of Figures 4a and 4b shows that for those MSTs with no routing errors, the lack of variation in bias in the approximate range > -1.5 θ < 1.5 was no longer observed—bias for both 1-3-3 and 1-3-4 MSTs tended to alternate from positive to negative across adjacent values of θ, with the effect more predominant for forward assembled (21, 14, 7) MSTs. Beyond that θ range, bias at both ends of the θ scale was very similar to that obtained when both error conditions were combined (Figure 4a).

The results of conditional mean bias grouped by the number of routing errors are shown in Figure 4c for 1-3-3 MSTs and Figure 4d for 1-3-4 MSTs, respectively (numerical values by MST condition are in Appendix Tables B and C) . The results show considerable bias in the s for MSTs across a relatively wide range of *θ,* especially for simulees with routing errors. 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 4a—the positive and negative estimation errors canceled each other, resulting in near zero bias when the biases of simulees with different numbers of misroutings were averaged. For the 1-3-3 design (Figure 4c), 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 4d), the same pattern was evident for the (21, 14, 7) MST but a more complex pattern emerged for MSTs with fewer items in the first stage. For both 1-3-3 and 1-3-4 MSTs the highest amount of bias occurred at extreme negative values of θ (> -1.5) with values as high as 1.35 for 1-3-3 designs (Table B) and 1.64 for 1-3-4 designs (Table C).

***F*igure 4. Mean conditional bias**

**a. All routing error conditions b. MSTs with zero routing errors**

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Diagram

Description automatically generatedDiagram

Description automatically generated**c. By number of routing errors d. By number of routing errors  
 for 1-3-3 MSTs for 1-3-4 MSTs**

**A picture containing table

Description automatically generated**

**RMSE**

Figure 5 shows the conditional RMSE for all conditions. MSTs performed the best near the center of the *θ* scale and poorest at the extremes, reflecting the information structure of the MST modules. 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 (Figure 5a). 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 for both 1-3-3 and 1-3-4 design, with the effect more obvious for 1-3-4 MSTs. There was no obvious effect of assembly priority. Figure 5b presents the conditional RMSE for MSTs with no routing errors, but there was little difference in RMSEs from the results for all MSTs regardless of the existence of routing errors (Figure 5a). As shown in Figures 5c and 5d, having more routing errors tended to produce higher RMSEs, especially near the center of the *θ* scale for simulees with two misroutings. RMSEs tended to be higher for negative values of θ (numerical values are in Tables D and E.

**Mean SEM**

The conditional mean SEM for all conditions is displayed in Figure 6. The MSTs exhibited a larger SEM toward the extreme ends of the θ scale, especially at the lower end. Differences among the MSTs were observed only for the extreme *θ*s. The 1-3-4 designs tended to have smaller SEMs than the 1-3-3 designs (Figure 6a). 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. Comparison of all simulees regardless of routing error status with those with no routing errors (Figures 6a vs. 6b) indicated little difference between the SEMS for the two groups. Misroutings did not have much effect on the SEMs for the 1-3-3 design (Figure 6c) but had a larger effect for the 1-3-4 design (Figure 6d), particularly for simulees that had two routing errors. For this latter group, the effect of routing errors for simulees with two routing errors was obvious for low *θ*s and varied differeently for forward and backward assembly as well as the distribution of items across the modules. As for bias and RMSE, values of SEM were highest at negative θ values (numerical values for all panels of Figure 6 are in Tables F – I).

**igure 5. Mean conditional RMSE**

Chart

Description automatically generatedShape

Description automatically generated with low confidence **a. All routing error conditions b. MSTs with zero routing errors**

**c. By number of routing errors d. By number of routing errors  
 for 1-3-3 MSTs for 1-3-4 MSTs**

A picture containing diagram

Description automatically generated**A picture containing table

Description automatically generated**Diagram

Description automatically generated

**Figure 6. Mean conditional SEM**

A picture containing shape

Description automatically generatedA picture containing engineering drawing

Description automatically generated **a. All routing error conditions b. MSTs with zero routing errors**

**c. By number of routing errors d. By number of routing errors  
 for 1-3-3 MSTs for 1-3-4 MSTs**

A picture containing diagram

Description automatically generatedA picture containing engineering drawing

Description automatically generated

**Discussion and Conclusions**

**Misrouting**

The results of the current study demonstrate that, averaging across θs and all investigated conditions, routing errors occurred 16.3% of the time, 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 between two θ anchors: 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). It should be noted that because this was a Monte Carlo simulation study, the responses of all simulees were fit to fit the model. But responses of real examinees are not always model fitting; therefore, it could be expected that with real examinees, percentages of misrouting could be even higher than that observed in this study.

Routing errors were also found to adversely affect bias, RMSE and, to a lesser extent, SEM. 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.

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 (Warm, 1989; 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**

In large and well-funded testing programs, MSTs can be assembled using algorithms that will attempt to optimize the allocation of items across the modules and panels to meet the requirements of a specific testing application; Zheng et al. (2014) provide an overview of these methods. However, for test developers who do not have the luxury of using these types of algorithms, the current study also provides some 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 also demonstrated thatplacing more items in the last stage tended to produce more precise measurement, supporting the rationale of adaptive testing that examinees should receive more items closely matched to their ability or trait level. 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 (i.e., 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. Since routing errors can only be fully identified by means of Monte Carlo simulations, MST developers should implement simulations prior to fielding an MST to evaluate the magnitude and effects of routing errors for candidate MST designs and select the design that minimizes those errors while having the least effect on measurement accuracy. At the same time, MST researchers should seek ways to reduce the number and effects of routing errors, since they are an intrinsic characteristic of MSTs.

**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 in some testing programs to ensure content coverage and test security. Future research can assess the impact of these practical constraints on the functioning of MSTs. 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*.* Future research shouldalso investigate alternative MST designs beyond the specific 1-3-3 and 1-3-4 designs used here. In addition, the results of this study are dependent on the routing rule used, in this case identify the module that provided maximum information at the simulee’s *θ* estimate at the routing juncture. Other routing rules are possible (e.g., Weissman, 2014) and should be investigated in future studies.

**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, although the present results indicate that the (21,14,7) design had the smallest percent of routing errors but also the poorest performance in terms of 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 and precision for examinees who true *θ*s deviate from the average *θ* for a group of examinees. MSTs might 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 potentially substantial proportions 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.

Han, K. (2020). Framework for developing multistage testing with intersectional routing for short-length Tests. *Applied Psychological Measurement, 44(2)*, 87–102

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.

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.

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). (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., 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.

Magis, D., Yan, D., & von Davier, A.(2018). mstR: Procedures to generate patterns under multistage testing. Available at https://cran.r-project.org/web/packages/mstR/index.html

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.

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.

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

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

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427-450.

[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.](https://journals.uair.arizona.edu/index.php/jmmss/article/view/15990) *Journal of Methods and Measurement in the Behavioral Sciences*, 2(1), 80-101.

Weissman, A. (2014). In D. Yan, A. A. von Davier, and C. Lewsis (Eds.). IRT-based multistage testing. In D. Yan, A. A. von Davier, and C. Lewis (Eds.). *Computerized multistage testing: Theory and applications.* Boca Raton FL: CRC Press.

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: Design, heuristic assembly, and comparison with other testing modes.* ACT Research Report Series, 2012 (6). Iowa City, IA: ACT.

Zheng, Y, Wang, C., Culbertson & Chang, H.-H. Overview of test assembly methods in multistage testing. (2014). In D. Yan, A. A. von Davier, and C. Lewis (Eds.). *Computerized multistage testing: Theory and applications.* Boca Raton FL: CRC Press.