**Effect of Routing Errors on the Psychometric Properties of Multistage Tests  
Robert Chapman, 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,” was 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 suggest 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

The primary objective of this study was to examine, in simulation, how routing errors (i.e., misrouting) 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

Four MST design factors (test structure, item allocation, assembly priority, and routing strategy) were manipulated. All MST designs were assembled from a “master” item bank. The items that were actually selected 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. Four MST design factors were manipulated to test which MST design yielded the best performance. This resulted in a total of 2 × 3 × 4 × 3 = 72 MST conditions. Table 1 shows how various conditions for test structure, assembly priority, and number of items were used to construct the MSTs.

**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 |
|  | Spiral | 7 | 14 | 21 |
|  |  | 14 | 14 | 14 |
|  |  | 21 | 14 | 7 |
|  | Random | 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 |
|  | Spiral | 7 | 14 | 21 |
|  |  | 14 | 14 | 14 |
|  |  | 21 | 14 | 7 |
|  | Random | 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 shown in Figure 1. 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, backward, spiral, and random). 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 & Chang, 2012). Two other assembly priority methods were used—the spiral assembly method, where assembly begins with the middle-most modules and ”spirals” outward (alternates forward and backward from the middle-most modules) and random assembly, where modules are assembled in random order (Zheng, Wang, Culbertson, & Chang, 2014).

*Routing Strategies.* The fourth factor was the method for making routing decisions between modules. Three routing strategies were evaluated from two broad categories of routing methods: appropriate maximum information (Luecht et al, 2006) and defined population intervals (Luecht, Brumfield, & Breithaupt, 2006; Zenisky, 2004). In the maximum information (MI) routing method, the module with the maximum information at a simulee’s incremental *θ* estimate is selected and administered. In defined population interval (PI) methods, the population distributions across the *θ* continuum (*θ* = −3:3) and number-correct spectrum (NC 0%:100%) are used to create sets of equal intervals. These intervals (*θ* or NC) are used to select and administer modules based on simulee incremental NC or estimated *θ*. Figure 2 displays the population distribution interval routing methods used in this study.

### 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

and , 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)] = 665 items, so 1,500 items were determined to be more than sufficient.

**Figure 2. *θ* and Number-Correct Routing Decision Methods**

**Table 2. Descriptive Statistics for Item Parameters**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Parameter | Mean | SD | Minimum | Maximum | Distribution |
| *a* | 1.28 | 0.21 | 0.95 | 2.20 | *ln*N(0.75, 0.25) |
| *b* | -0.01 | 0.80 | -1.83 | 1.5 | N(0, 1) |
| *c* | -0.15 | 0.03 | 0.10 | 0.2 | 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 appropriate 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, backward, spiral, or randomly assembled. Figure S-1 (in the online supplement) 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.

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 according to either the θ or NC routing methods. 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 along a pathway that did not match the most common pathway for all simulees of a given true *θ.* Routing errors resulting from transitions between stages 1 and 2 as well as stages 2 and 3 were analyzed.

**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,* observedpath (including misrouted simulees), and observed path by *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 8.2% for a randomly assembled MST that used MI routing with half the test items (21) in the first stage to 25.5% for spiral assembled MSTs that used MI routing with 7 items in the first stage. The average number of routing errors across all conditions was 15% (9.2% were misrouted with a single path error and an additional 5.8% were misrouted with two path errors). Overall, 1-3-3 MSTs had a lower average percentage of routing errors (8.2% to 19.1%) than 1-3-4 MSTs (10.2% to 25.5%), likely because there were more routing targets in the third stage of 1-3-4 MSTs. Across all conditions, allocating most items in the first stage also resulted in a lower percentage of routing errors (8.2% and 19.7% across all “decreasing” designs) than allocating most items in the last stage (15.3% and 25.5% for “increasing” designs). Different assembly methods did not result in substantial differences in the range of misrouting errors across conditions, with the smallest range in routing errors being less than 1% (ranging from 13.20% to 14.10% for random and backward assembly of a 1-3-3 test with an equal number of items using MI routing) and the largest range of routing errors being 3% (ranging from 8.2% to 11.2% for random and spiral assembly of a 1-3-3 test with a decreasing number of items and MI routing). MI routing resulted in the largest range of routing errors (8.2% to 25.5%) as compared to PI routing methods (*θ* cutscore, 8.8% to 20.3%; NC cutscore 8.7%- to 9.1%).

**Table 3. Percentage of Routing Errors for Each MST Design**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Assembly Design | | | | | | | |
|  | Forward | | Backward | | Spiral | | Random | |
|  | a. MI Routing | | | | | | | |
| Number of items | 1-3-3 | 1-3-4 | 1-3-3 | 1-3-4 | 1-3-3 | 1-3-4 | 1-3-3 | 1-3-4 |
| Increasing | 17.2% | 25.1% | 16.9% | 23.4% | 15.7% | 25.5% | 17.1% | 23.6% |
| Equal | 13.8% | 21.2% | 14.1% | 20.2% | 13.6% | 20.0% | 13.2% | 18.5% |
| Decreasing | 9.8% | 18.8% | 10.4% | 19.5% | 11.2% | 19.7% | 8.2% | 19.2% |
|  |  |  |  |  |  |  |  |  |
|  | b. PI-*θ* Routing | | | | | | | |
|  | 1-3-3 | 1-3-4 | 1-3-3 | 1-3-4 | 1-3-3 | 1-3-4 | 1-3-3 | 1-3-4 |
| Increasing | 17.2% | 19.4% | 18.4% | 19.1% | 16.9% | 20.3% | 16.8% | 17.9% |
| Equal | 10.6% | 15.1% | 12.5% | 16.4% | 11.1% | 15.5% | 11.3% | 14.3% |
| Decreasing | 8.8% | 14.0% | 8.8% | 13.6% | 9.8% | 15.0% | 9.3% | 13.7% |
|  |  |  |  |  |  |  |  |  |
|  | c. PI-NC Routing | | | | | | | |
|  | 1-3-3 | 1-3-4 | 1-3-3 | 1-3-4 | 1-3-3 | 1-3-4 | 1-3-3 | 1-3-4 |
| Increasing | 19.1% | 15.3% | 18.8% | 16.5% | 17.5% | 17.5% | 18.0% | 15.7% |
| Equal | 12.2% | 13.3% | 11.9% | 11.2% | 11.7% | 12.1% | 12.9% | 12.6% |
| Decreasing | 9.1% | 11.0% | 9.6% | 10.9% | 8.7% | 11.6% | 9.2% | 10.2% |

**-**

Figure S-2 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 of a module (e.g., *θ* = −1, *θ* = 0 and *θ* = 1 with 1-3-3 designs) the percentage of routing errors was relatively low. But when the simulee’s *θ* was between two *θ* anchors of adjacent modules, the percentage of routing errors increased. Comparing the 1-3-3 and 1-3-4 designs, the 1-3-4 designs (Figures S-2b, S-2d, S-2f) generally misrouted simulees over a wider range of *θ*. However, the number of routing errors decreased with increasing numbers of items in the initial module for the 1-3-3 design (Figures S-2a, S-2c, S-2e) but not for the 1-3-4 design (Figures S-2b, S-2d, S-2f). 1-3-3 MSTs also had more simulees who had two routing errors than did the 1-3-4 MSTs. There were not regular or pronounced differences in the distribution of routing errors specific to the assembly methods. The distribution and magnitude of the routing errors was very similar between the MI and PI-*θ* cutscore routing methods using a 1-3-3 design (Figures S-2a, S-2c) but differed slightly in the 1-3-4 design (Figures S-2b, S-2d) where the PI methods resulted in errors more narrowly concentrated around the θ cut points (Figure S-2c through S-f). 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.

**Overall Mean Bias, RMSE and SEM by Condition and Number of Path Errors**

Table 4 shows overall bias, RMSE, and mean SEM in each manipulated condition by number of path errors. Overall bias was generally low across conditions, with the greatest bias occurring for simulees with zero errors (bias = −0.176) and simulees whose tests used a decreasing number of items across stages (bias = −0.103). Overall, RMSEs and mean SEMs varied across conditions but were higher for the group of simulees with zero errors (RMSE 0.454 to 0.724 and mean SEM 0.701 to 1.244) compared to the groups of simulees who experienced either one or two errors (RMSE 0.204 to 0.387 and mean SEM 0.191 to 0.483). This discrepancy in the magnitude of RMSE and mean SEM results from the general pattern of the path errors shown in Figure S-2, where errors tended to occur when simulee true *θ*s were close to module information centers (Figure 1). In other words, the subset of simulees with one or two path errors underrepresents the portion of simulees drawn from a uniform distribution with extreme *θ* levels (e.g., *θ* > 2 or *θ* < −2), which limits the magnitude of the observed RMSEs and mean SEMs. These simulees are likely to have high SEMs (e.g., > 1) due to large mismatches between item difficulty and simulee true *θ,* and atthe extremes of true *θ* simulees may also have non-mixed response vectors and greatly inflated SEMs (SEM > 4). Across test structures, both 1-3-3 and 1-3-4 tests showed low bias (bias = −0.086 to −0.005). Among simulees that had only one routing error, the 1-3-4 tests demonstrated lower RMSEs and mean SEMs than the 1-3-3 tests (0 errors, RMSE = 0.555 vs. 0.569 and mean SEM = 0.867 vs 0.991; 1 error, RMSE = 0.204 vs 0.281 and mean SEM = 0.191 vs. 0.292), with the opposite trend (higher RMSEs and mean SEMs) occurring in the 1-3-4 test with simulees who had two routing errors (RMSE 0.387 vs. 0.251 and mean SEM 0.444 vs. 0.207). This is expected due to the broader range of ability content included in the final stage of the 1-3-4 test modules compared to the 1-3-3 test modules (see Figure 1). Across number of items, simulees who had zero errors showed low bias (bias = −0.056 to −0.103).

Among simulees with zero errors, increasing item allocation from earlier stages to later stages resulted in greater measurement precision as indicated by comparatively lower RMSEs and mean SEMs (RMSE 0.603 to 0.501 and mean SEM 1.124 to 0.701). Among simulees who had one or more path errors, the opposite trend occurred, with higher RMSEs and mean SEMs in tests that allocated more items to later stages (one error, RMSE 0.209 to 0.267 and mean SEM 0.203 to 0.272; two errors, RMSE 0.282 to 0.365 and mean SEMs 0.217 to 0.456). Across assembly methods, simulees with zero errors showed low bias (bias = −0.09 to −0.081) and very similar RMSEs and mean SEMs. Across routing methods, the MI and PI-*θ* cutscore methods similarly outperformed the PI-NC routing method (0 errors, bias −0.041 to −0.034 vs. −0.176 and RMSE 0.46 to 0.454 vs. 0.724 and mean SEM 0.772 to 766 vs. 1.244; one error, RMSE 0.221 to 0.213 vs. 0.318 and mean SEM 0.214 to 0.223 vs. 0.298). Among the simulees with two path errors, the MI method was outperformed by the PI methods in terms of RMSE and mean SEM (RMSE 0.38 vs. 0.317 to 0.315) and mean SEM 0.483 to 0.263 to 0.300.

**Overall Mean Bias, RMSE and SEM by Path**

Tables S-1 and S-2 show the overall bias, RMSE, SEM, and percent of errors in each manipulated condition by path (see Figure 3 for module numbers) in the 1-3-3 and 1-3-4 tests. Bias, RMSE, and mean SEM generally increased as paths aligned with more extreme *θ* levels (e.g., paths 1-2-5 and 1-4-7 or 1-4-8). This makes sense given that simulees with extreme levels of ability that exceed the information centers of the modules (Figure 1) are correctly routed to these paths. This trend is also reflected in the lower percentages of path errors among those simulees who were administered the upper-most (1-2-5) or lower-most (1-4-7 or 1-4-8) path (0-3% in the 1-3-3 test and 2% to 6% in the 1-3-4 test). In the 1-3-3 test, the off-center paths with at least one routing error (e.g., 1-3-5 and 1-2-6) resulted in 100% error and the center path (1-3-6) demonstrated a substantial error rate (39% to 45%). In the 1-3-4 test, two center paths exist (1-3-6 and 1-3-7) and a trend similar to what was found in the 1-3-3 test emerged for off-center paths (e.g., 1-2-6 and 1-4-7). These paths generally resulted in higher error percentages (29% to 100%) than center paths (35% to 69%) in the 1-3-4 test. Interestingly, the off-center paths (1-2-6 and 1-4-7) tended to have lower error statistics (bias, RMSE, mean SEM) than the center paths (1-3-6 and 1-3-7). There was no noticeable effect of assembly method on error statistics by path. With

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Test Structure | | Number of Items by Module | | | Assembly Method | | | | Routing Method | | |
| Number of Path Errors | Statistic | 1-3-3 | 1-3-4 | Increasing | Equal | Decreasing | Forward | Backward | Spiral | Random | Maximum Information | *θ*  Score | Number-correct |
| 0 errors | Bias | -0.084 | -0.086 | -0.056 | -0.094 | -0.103 | -0.081 | -0.083 | -0.085 | -0.09 | -0.041 | -0.034 | -0.176 |
| RMSE | 0.569 | 0.555 | 0.501 | 0.574 | 0.603 | 0.562 | 0.562 | 0.561 | 0.563 | 0.46 | 0.454 | 0.724 |
| SEM | 0.991 | 0.867 | 0.701 | 0.949 | 1.124 | 0.919 | 0.943 | 0.917 | 0.943 | 0.772 | 0.766 | 1.244 |
| 1 error | Bias | -0.007 | -0.005 | -0.009 | -0.004 | -0.004 | -0.011 | -0.014 | -0.008 | 0.009 | -0.007 | -0.011 | 0.005 |
| RMSE | 0.281 | 0.204 | 0.267 | 0.236 | 0.209 | 0.245 | 0.226 | 0.249 | 0.243 | 0.221 | 0.213 | 0.318 |
| SEM | 0.292 | 0.191 | 0.272 | 0.22 | 0.203 | 0.225 | 0.228 | 0.242 | 0.246 | 0.223 | 0.214 | 0.298 |
| 2 errors | Bias | -0.017 | -0.031 | -0.024 | -0.031 | -0.019 | -0.023 | -0.025 | -0.024 | -0.026 | -0.038 | -0.048 | -0.002 |
| RMSE | 0.251 | 0.387 | 0.365 | 0.324 | 0.282 | 0.318 | 0.318 | 0.338 | 0.366 | 0.38 | 0.317 | 0.315 |
| SEM | 0.207 | 0.444 | 0.456 | 0.259 | 0.217 | 0.311 | 0.313 | 0.325 | 0.425 | 0.483 | 0.263 | 0.300 |

**Table 4. Overall Bias, RMSE, and Mean SEM, by Condition and Number of Path Errors**

**Figure 3. Module Numbers by Design Structure**

**a. 1-3-3 Design b. 1-3-4 Design**

increasing numbers of items allocated to later stages of the test, all error statistics generally improved. Across routing methods, there was generally higher error statistics associated with the 1-2-5 path for the maximum information and population interval distribution *θ* cutscore methods and conversely higher error associated with the 1-4-7 and 1-4-8 paths for the PI-NC cutscore routing method.

**Mean Bias Conditional on *θ***

Figure S-3 shows mean bias across *θ* for each MST design averaged across routing error conditions. MSTs of all conditions performed equally well near the center of the *θ* scale for *θ* between −1.0 and 1.0, 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— the exception being the tests that were administered using the PI-NC cutscore (Figures S-3e and S-3f), which showed substantial error at lower positive levels of *θ* due to simulee guessing influencing the NC scores. On the positive end of θ, higher levels of bias occurred at *θ* = 1.5 and above for MI and PI-*θ* cutscores (Figure S-3a, S-3b, S-3c, S-3d), with the values of bias in that region generally higher than those observed for negative *θ*s. The underestimation and overestimation were larger for 1-3-3 MSTs (Figures S-3a, S-3c, S-3e) than 1-3-4 MSTs (Figures S-3b, S-3d, S-3f), also for MSTs that had more items in the earlier stages than MSTs that had more items in the later stages (“increasing”). There was no general effect of assembly priority.

The results of conditional mean bias grouped by the number of routing errors are shown in Figure S-4. 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 (solid lines). Simulees with one or two misroutings generally displayed the opposite trend (dashed lines), noting some instability in the estimates due to the previously noted small number of misrouting errors at more extreme levels of *θ* . This explains the zero bias near the center of the *θ* scales in Figure S-3—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 both the 1-3-3 and 1-3-4 designs using MI and PI-*θ* routing (Figure S-4a, S-4b, S-4c, S-4d), bias for simulees with two misroutings (dotted line) was generally higher than for those with a single misrouting (dashed line), but was concentrated more toward the center of the *θ* scale. For the test that used the PI-NC routing (Figure S-4e, S-4f), the same pattern was not evident, possibly due to the small number of routing errors in the extremes for the 1-3-3 and 1-3-4 (Figures S-4e and S-4f) for simulees with single routing errors (dashed line).

**RMSE Conditional on *θ***

Figure S-5 shows the conditional RMSE for all manipulated 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 for the MI routing (Figures S-5b vs. S-5a) and PI-*θ* routing and (Figures S-5d vs. S-5c). Additional items in the last stage (“increasing”) led to smaller RMSEs at the extremes for both 1-3-3 and 1-3-4 designs, with a more pronounced effect for 1-3-4 MSTs that used MI and PI-*θ* routing (Figures S-5b, S-5d). There was no obvious effect of assembly priority. In the MI and PI-*θ* routing methods (Figures S-5a, S-5b, S-5c, S-5d), having more routing errors tended to produce higher RMSEs, especially near the center of the *θ* scale for simulees with two misroutings (dotted line). RMSEs tended to be higher for positive extreme values of *θ* across conditions. Overall, the magnitude and shape of RMSEs were similar between the MI and PI-*θ* routing methods (Figures S-5a, S-5b, S-5c, S-5d). In comparison, PI-NC routing (Figures S-5e, S-5f) resulted in similar patterns but overall higher RMSEs.

**Mean SEM Conditional on *θ***

The conditional mean SEM for all conditions is displayed in Figure S-6. The MSTs exhibited a larger SEM toward the extreme ends of the *θ* scale, especially at the lower end for MI and PI-*θ* routing and the higher end for PI-NC routing. Differences among the MSTs were observed only for the extreme *θ*s. The 1-3-4 designs (Figures S-6b, S-6d, S-6f) tended to have smaller SEMs than the 1-3-3 designs (Figures S-6a, S-6c, S-6e). Having more items allocated to later stages (“increasing”) resulted in smaller SEMs. Assembly method did not generally impact SEMs. Comparison of simulees without routing errors (solid lines) to those with routing errors (dashed or dotted lines) indicated little difference between the groups, noting that few routing errors existed at extreme *θ* levels, which sometimes resulted in unstable mean SEM estimates. Across routing methods, the MI and PI methods resulted in very similar magnitudes and pattern of SEMs (Figures S-6a, S-6b, S-6c, S-6d), with higher SEMs at lower extreme levels of *θ* . In comparison, PI-NC routing (Figures S-6e, S-6f) resulted in the opposite pattern where higher SEMs were located at higher extreme levels of *θ* .

**Discussion and Conclusions**

**Misrouting**

The results of the current study demonstrate that, averaging across θs and all investigated conditions, routing errors occurred 15% of the time, with a range of 8.2% to 25.5%, depending on item allocation in the MSTs. Placing more items in the first stage of an MST (“decreasing”) resulted in the fewest routing errors, whereas placing more items in the last stage test (“increasing”) 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 probabilistically generated based on an IRT 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 and *θ* levels. Misrouted simulees with *θ*s near the anchor points had high magnitudes of bias (over-estimation at low levels of *θ* and under-estimation at high levels of *θ* ) 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 between simulees with and without misroutings. This indifference contrasts sharply with the results obtained for bias and RMSE and suggests that the SEM of the *θ* estimates as computed from Equation 5 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. The 1-3-3 designs resulted in fewer misroutings than for 1-3-4 designs, likely due to fewer routing options. The smallest percentage of misroutings occurred for a randomly assembled 1-3-3 MST with 21 items in Stage 1 and the largest (by more than three times) for a spiral assembled 1-3-4 design with seven items in the first stage.

When examining the researcher choice of how to make routing decisions, there were two broad strategies evaluated: the maximum information approach which leverages IRT-based information from test modules and individual examinees to make routing decisions, or the population interval distribution method which utilizes known population score distributions of IRT or number-correct scores to create cut point thresholds for routing. Across all metrics, the population interval method using number-correct routing was outperformed by routing methods that utilized elements of IRT (i.e., maximum information and population interval distribution θ cutscore routing methods), with one exception. In the case of simulees with two path errors, population interval distribution routing appeared to do better than maximum information routing. However, this may be an artifact of where routing errors occur on the *θ* spectrum; simulees with two errors who are routed with population interval distribution cut points are concentrated near the center of measurement space (*θ* = −1 to 1) where greater item information and better measurement precision exists. In contrast, simulees with two errors who are routed with the maximum information approach have a broader and more diffuse span across the *θ* spectrum, perhaps reflecting the reality that examinee non-model fitting responses may exist at all levels of ability. Another concern with employing the population interval distribution cutscores is the assumption of the population distribution of scores. It is unlikely that the test designer knows the underlying population score distribution with certainty. Although not examined in this study, it is logical to assume that a mismatch between the choice of population-derived score cut points and the realized empirical distribution of scores would lead to additional error, both in terms of path routing and measurement precision.

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 beyond that which is inherent in the form of MSTs was implemented in the simulations or manipulated. In practice, these additional constraints are often imposed in some testing programs to ensure content coverage and test security, but in contrast to item-by-item CAT these are usually integrated into the module design stages in MSTs. Nevertheless, 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 whose 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 tests 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). 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.

Zenisky, A. L. Evaluating the effects of several multi-stage testing design variables on selected psychometric outcomes for certification and licensure assessment. *University of Massachusetts Amherst*; 2004.

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, M. & 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.