## Purpose of the Study

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 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.

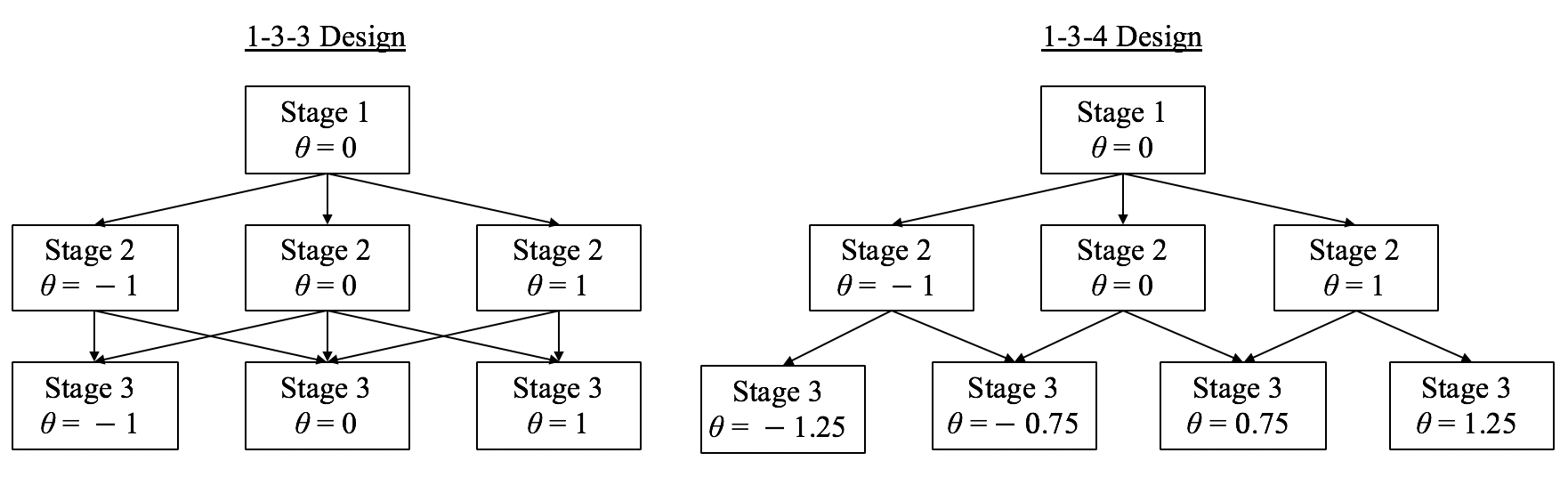
Conditions. Three MST design factors were manipulated to test which MST design yielded the best performance. This resulted in a total of 2 × 3 × 4 × 3 = 72 MST conditions. Table 1 shows how various conditions for test structure, assembly priority and number of items were used to construct the MSTs evaluated in this study.

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 |
|  | 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 |
|  |  |  |  |  |

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).

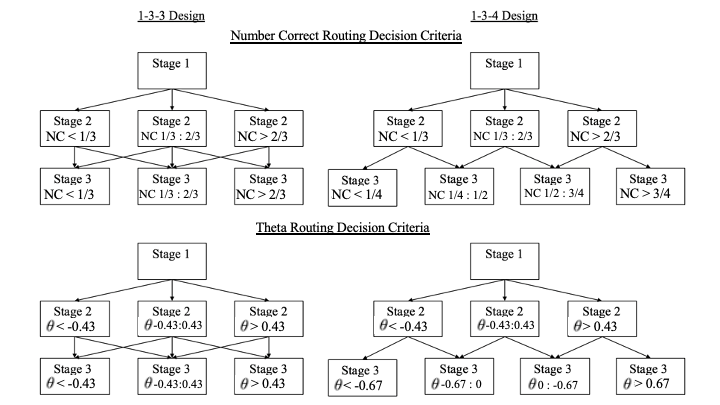
**Figure 1*.* 1-3-3 and 1-3-4 MST designs** **

### 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, 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 and 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, Yi & Wang, Chun & Culbertson, Michael & Chang, Hua-Hua, 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, and Breithaupt, 2006; Zenisky, 2004). In the maximum information routing method, the module with the maximum information at simulee’s incremental theta estimate is selected. In defined population interval methods, the population distributions across the theta continuum (Theta -3:3) and number correct spectrum (NC 0%:100%) are used to create sets of equal intervals. These intervals (theta or NC) are used to select modules based on simulee incremental NC or estimated theta. Figure X. represents the population distribution interval routing methods used in this study.

Figure X. Theta and Number Correct Routing Decision Methods



### 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, backward, spiral or randomly 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 according to the appropriate maximum information or population distribution interval (theta 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 to a module that did not match the true *θ* or final NC, or 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 than they 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,* 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.

**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.

Zenisky AL. 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 & 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.