Listening. Learning. Leading.®

Research Report
ETS RR–14-01

# An Investigation of the Impact of Misrouting Under Two-Stage Multistage Testing: A Simulation Study

Sooyeon Kim

Tim Moses

June 2014

Discover this journal online at
Wiley Online Library
wileyonlinelibrary.com

# ETS Research Report Series

## EIGNOR EXECUTIVE EDITOR

James Carlson
*Principal Psychometrician*

## ASSOCIATE EDITORS

Beata Beigman Klebanov
*Research Scientist*

Heather Buzick
*Research Scientist*

Brent Bridgeman
*Distinguished Presidential Appointee*

Keelan Evanini
*Managing Research Scientist*

Marna Golub-Smith
*Principal Psychometrician*

Shelby Haberman
*Distinguished Presidential Appointee*

Gary Ockey
*Research Scientist*

Donald Powers
*Managing Principal Research Scientist*

Gautam Puhan
*Senior Psychometrician*

John Sabatini
*Managing Principal Research Scientist*

Matthias von Davier
*Director, Research*

Rebecca Zwick
*Distinguished Presidential Appointee*

## PRODUCTION EDITORS

Kim Fryer
*Manager, Editing Services*

Ayleen Stellhorn
*Editor*

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

RESEARCH REPORT

# An Investigation of the Impact of Misrouting Under Two-Stage Multistage Testing: A Simulation Study

Sooyeon Kim & Tim Moses

Educational Testing Service, Princeton, NJ

The purpose of this study was to investigate the potential impact of misrouting under a 2-stage multistage test (MST) design, which includes 1 routing and 3 second-stage modules. Simulations were used to create a situation in which a large group of examinees took each of the 3 possible MST paths (high, middle, and low). We compared differences in examinees' scores associated with different paths under 2 MST assembly conditions: a small-difference condition in which 3 modules at Stage 2 overlap in difficulty, and a large-difference condition in which the 3 modules at Stage 2 are distinct in difficulty. We also compared examinees' score differences associated with the target MST path with the ones obtained when the second-stage module was 1 level of difficulty above or below the target module. When the second-stage modules overlapped in difficulty, the score differences (i.e., bias) associated with different MST paths were negligible for practical purposes. Similar trends appeared even when the second-stage modules were significantly distinct in difficulty. The impact of misrouting was generally minimal under the MST design used in this study.

**Keywords** Multistage testing; routing; panel assembly; adaptive test; bias

doi:10.1002/ets2.12000

## Multistage Testing

Multistage testing (MST) is a procedure designed to provide the benefits of adaptive testing—improved measurement precision and efficiency—without the problems that arise from testing each examinee with a different set of test items. Rather than custom building a test form for each examinee as the test unfolds, MST routes examinees through a series of preassembled test modules. The testing procedure is divided into stages. In the first stage, there is only one module; all examinees are tested with the same set of items. In the second stage, there are two or more modules that differ systematically in difficulty. Each examinee is assigned to a second-stage module on the basis of his or her performance at the first stage. If there is a third stage, it is adaptive in the same way as the second stage.

Although a computerized adaptive test (CAT) and an MST are both adaptive, the testing designs differ substantially in their adaptive algorithms. Under the CAT design, item selection algorithms construct each test form while the examinee is taking the test by iteratively administering an item, estimating a provisional score, and then selecting the next item from the active item bank using certain statistical optimization criteria (Luecht & Nungester, 1998). Therefore, adaptation to an examinee's ability takes place at the item level. Under the MST design, adaptation to an examinee's ability occurs between stages of the testing process and is based on the examinee's cumulative performance on previous item sets. Accordingly, adaptation to an examinee's ability takes place between stages and, thus, fewer adaptation points are available under MST.

Examinees under CAT must answer each item when it is presented; they cannot skip ahead or go back and change an answer. Examinees under MST can answer the items in any order and can review or change their answers within a particular module. Several surveys have indicated that examinees, particularly strong examinees, prefer being able to navigate through an examination section following their own sequence of items (Melican, Breithaupt, & Zhang, 2010; Parshall, Spray, Kalohn, & Davey, 2002; Wise, 1996). Because an MST consists of a small number of separate modules, the test developers can make sure each module meets a set of specifications for item content, item difficulty, total word count, and distribution of answer key positions. They can also avoid dependencies between items in the same module. To achieve a given level of measurement precision requires administering more items to each examinee under MST than under CAT, but not as many as under conventional, nonadaptive testing. Because MST provides a balanced compromise

*Corresponding author:* Sooyeon Kim, E-mail: skim@ets.org

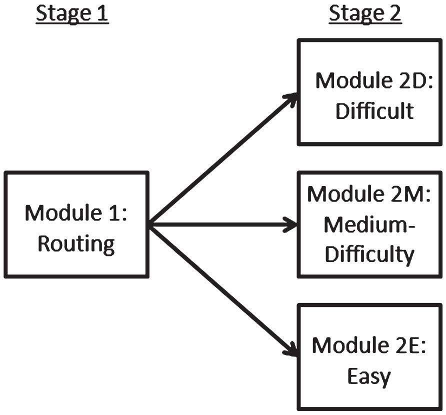

**Figure 1** Schematic of a two-stage multistage test.

between fully adaptive testing and nonadaptive testing, this compromise has led to many studies on MST (see Luecht & Sireci, 2011; Zenisky, Hambleton, & Luecht, 2010) and its operational use in practice (see Educational Testing Service, 2011).

The MST designs can vary, depending on score usage, psychometric characteristics, and administration environments. As for the number of stages and modules, several studies indicate that adding more stages and more modules in stages can produce tremendous complexity in practice without adding any further psychometric benefits on the final forms (Jodoin, Zenisky, & Hambleton, 2006; Luecht & Nungester, 1998; Luecht, Nungester, & Hadidi, 1996; Wang, Fluegge, & Luecht, 2012). A recent study (Wang et al., 2012) showed that both complex and simple MST design formats performed equally well when the item bank is optimal, with high quality items targeting key ability regions. In this study, the authors recommend the use of simple MST configurations such as 1–2 [Stage 1 (routing) – Stage 2 (high/low)] or 1–2–2 [Stage 1 (routing) – Stage 2 (high/low) – Stage 3 (high/low)], because those designs demand a modestly sized item bank. Kim and Plake (1993) found that the statistical characteristics of the first-stage (routing) module had a major influence on the complete test's measurement precision compared to the later stages' modules. Increasing the length of the routing module was most important in reducing the size of the proficiency estimate errors. This strategy may lead to some item exposure concern, however, because the routing module has to be administered to all examinees who take a particular MST.

A two-stage MST represents the simplest design of a multistage adaptive test. Figure 1 illustrates an example of a two-stage MST. Each module at Stage 2 concentrates on a particular level of difficulty to differentiate examinees' abilities within a certain range of proficiency after routing. The two-stage MST procedures are as follows. At Stage 1, examinees take a predetermined routing module. The examinees' ability will be initially estimated based on their performance on the routing module. To estimate the examinees' initial ability, the routing module generally includes items with a broad range of difficulty (easy to difficult). At Stage 2, three predetermined modules are administered. Each module concentrates on a particular level of difficulty, either low (easy), middle (medium-difficulty), or high (difficult). Examinees receive a set of items determined by their performance at Stage 1. For example, an examinee with a high initial ability estimate will receive a module consisting of difficult items at Stage 2, whereas a less capable examinee with a low initial ability estimate will receive a module consisting of less difficult items.

The two-stage MST is actually a sequence of two conventional linear tests, with the first test scored before administration of the second test and used to assign each examinee to the appropriate difficulty level of the second test. Although the two-stage MST has many advantages, it also has the potential disadvantage of a higher likelihood of routing error caused by fewer adaptation points compared to CAT. This likelihood is especially high for examinees whose scores fall near the routing cutscores. Such uncertainty may be exacerbated if too few items are available for routing and examinees guess item answers (Hendrickson, 2007). A further issue involves the capacity of the modules that follow at Stage 2 to estimate the examinees' actual ability adequately, despite the incorrect routing. This may depend on how much overlap in difficulty there is between the second-stage modules. For example, the effects of misrouting will be substantial if the second-stage modules focus only on a very narrow range of proficiency. Misrouting will be less of a problem if there is more overlap in item difficulty between the second-stage modules. The risk of misrouting could be crucial, particularly with a two-stage MST, because the opportunity for adaptation is limited. The choice of MST designs should be made after investigating any potential impact of misrouting, such as bias in examinees' ability estimates, caused by providing modules that are not matched with their ability levels.

## Purpose

The purpose of this study was to investigate the potential impact of misrouting under a two-stage MST, with one routing module at Stage 1 and three modules, differing systematically in difficulty, at Stage 2. It is hard to obtain data from a group of actual examinees who have completed all the paths within an MST without changing their abilities in the process. Thus, simulations were used to create a situation in which a large group of examinees took all three second-stage models with no change in their abilities. We compared differences in examinees' scores associated with different paths under two MST assembly conditions: a small-difference condition in which three modules at Stage 2 overlap in difficulty, and a large-difference condition in which the three modules at Stage 2 are substantially distinct in difficulty. The mean difference between any two paths should be close to zero (i.e., no bias) in a situation in which the impact of misrouting is trivial. Conversely, if the average difference between paths is substantial relative to a score scale (e.g., greater than half of a score scale point), the impact of misrouting would be considered nontrivial. We also compared examinees' scores associated with the target MST module, contingent upon their performance at Stage 1, with the scores obtained when the second-stage module was one level of difficulty above or below the target module (i.e., off level caused by misrouting). We examined the magnitude of bias and score variability caused by misrouting under two MST assembly conditions to determine whether the use of highly distinct Stage 2 modules is sensitive to misrouting.

# Method

## Multistage Test

We selected a two-stage MST design illustrated in Figure 1.[1] We use the term *path* to mean a combination of modules that possibly could be presented to an examinee. Each path consists of the first-stage module and one of the second-stage modules. There are three possible paths of the MST in Figure 1:

- High path: Module 1 (routing), Module 2D (difficult)
- Middle path: Module 1 (routing), Module 2M (medium-difficulty)
- Low path: Module 1 (routing), Module 2E (easy)

We simulated two MST panels based on the two assembly conditions. Each panel includes four modules, three paths, and 80 items (20 items per module). Item parameters for each of four modules on a particular MST panel were generated based on the two-parameter logistic (2PL) item response theory (IRT) model using the random number generator function built in Excel.[2] Table 1 summarizes the descriptive properties of item discrimination (a) and item difficulty (b) on each of the four modules under two MST assembly conditions (small vs. large). The Stage 2 easy module was much easier in the large-difference condition than in the small-difference condition. To create an easy

**Table 1** Descriptive Properties of Item Discrimination (a) and Item Difficulty (b) for Each of Four Modules Under Small- and Large-Difficulty Difference Conditions

| | | | *a* parameters | | *b* parameters | |
|---|---|---|---|---|---|---|
| Difficulty difference | Stage | Class/level | M | SD | M | SD |
| Small | 1 | Routing | 0.85 | 0.27 | −0.04 | 0.89 |
| | 2 | Difficult | 0.85 | 0.30 | 0.72 | 0.70 |
| | 2 | Medium | 0.85 | 0.29 | 0.00 | 0.69 |
| | 2 | Easy | 0.86 | 0.30 | −0.76 | 0.71 |
| Large | 1 | Routing | 0.85 | 0.27 | −0.04 | 0.89 |
| | 2 | Difficult | 0.85 | 0.30 | 1.42 | 0.70 |
| | 2 | Medium | 0.85 | 0.29 | 0.00 | 0.69 |
| | 2 | Easy | 0.86 | 0.30 | −1.46 | 0.71 |

*Note:* The IRT model based reliability was .93 for both small- and large-difference MSTs. The raw score SEM was 1.82 for the small-difference MST and 1.80 for the large-difference MST.

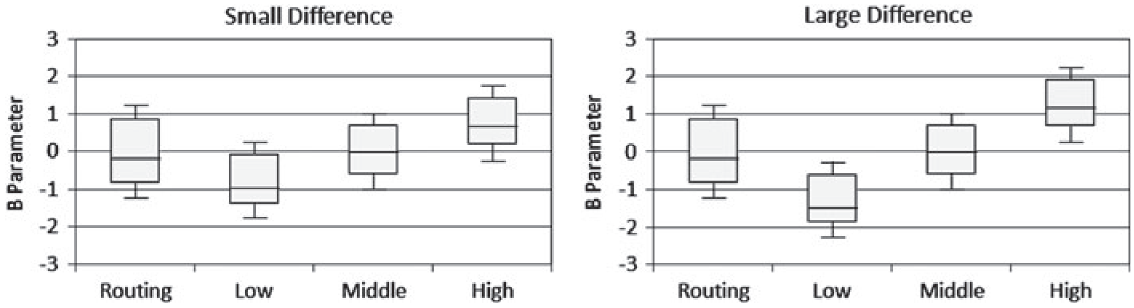

**Figure 2** Box-whiskers plots for each of four modules under small- and large-difficulty difference conditions.

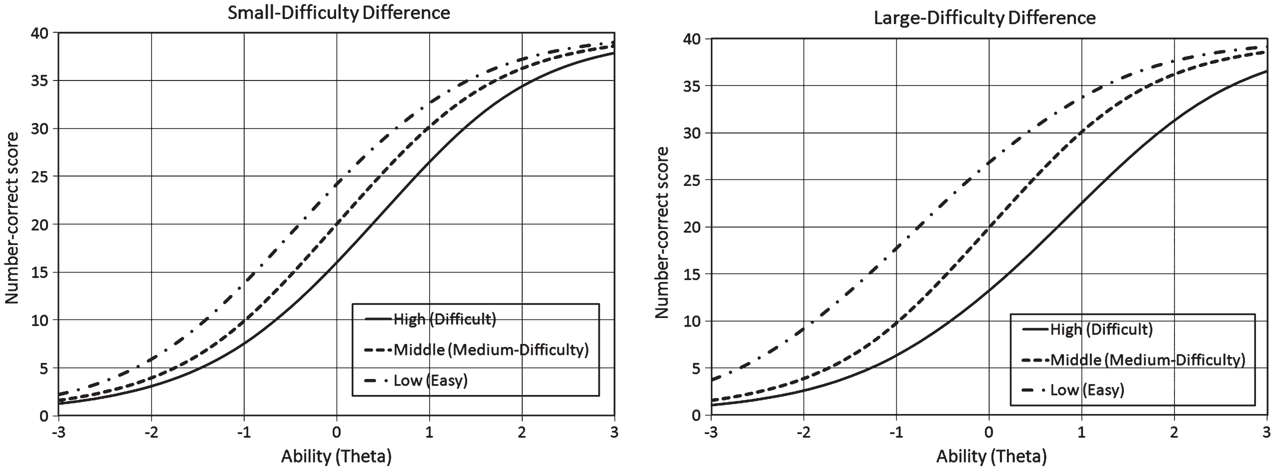

**Figure 3** The test characteristic curve for each of the three paths under the small- and large-difficulty difference conditions.

module in the large-difference condition, 0.70 (1 SD unit) was subtracted from the item difficulty parameters in the easy module of the small-difference condition. Conversely, the Stage 2 difficult module was much harder in the large-difference condition than in the small-difference condition. A difficult module in the large-difference condition was created by adding 0.70 (about 1 SD unit) to the item difficulty parameters in the difficult module of the small-difference condition. The routing module and Stage 2 medium-difficulty module were the same in both assembly conditions. Figure 2 graphically presents the difficulty differences between the two assembly conditions using box-and-whiskers plots.

There is more than one way to score an MST. We used number-correct scoring (i.e., summed scoring) in this simulation. Accordingly, the raw score on any of the three paths is simply the total number of items answered correctly in the two stages. The two plots in Figure 3 indicate the test characteristic curve (TCC) for each path under the small- and large-difference conditions, respectively. The total number-correct scores obtained from each of the three paths were transformed onto a common score scale using a hypothetical reference form that included a distribution of item difficulties across the entire theta scale.[3] Therefore, the raw-to-equated raw score conversion for each path of each MST panel was determined by IRT true-score equating, using the 2PL model.

## Procedures

We simulated 1,000 examinees at each of 41 quadrature points on a theta scale ranging from −3.0 (minimum) to +3.0 (maximum), with an interval of 0.15. Simulated examinees' thetas were uniformly distributed ($N = 41,000$). We used SAS statistical software to generate each simulated examinee's dichotomous response (correct or incorrect) to each item of the MST panel.

Target MST indicates the case where the simulated examinees receive the most appropriate second-stage module in light of their performance at Stage 1. In this simulation, we compared the target MST case with other off-level module cases.

On the basis of the *defined population intervals* method (Luecht, Brumfield, & Breithaupt, 2006), we selected two cutscores for routing simulated examinees to a target module so as to result in approximately 30%, 40%, and 30% of the examinees taking high, middle, and low paths, respectively, in a situation where the simulated examinees' ability distribution follows the standard normal distribution, that is, $\theta$ $N$ (0, 1).[4] The same cutscores were applied to both small- and large-difference conditions.[5]

The simulation procedure for each examinee, for each of the modules (in a small-difference condition), consisted of the following steps:

1. Generate the simulated examinee's response to each item in the Stage 1 module.
2. Assign the examinee to all three Stage 2 modules.
3. Generate the examinee's response to each item in the Stage 2 modules.
4. Compute the examinee's number-correct raw score on Stage 1 and Stage 2 and the total raw score for each of the three paths.
5. Apply the appropriate raw-to-equated raw score conversion obtained through IRT true-score equating to determine the examinee's equated raw score for each path. Name the scores as follows: *high* for high path, *middle* for middle path, and *low* for low path.
6. Apply the Stage 1 cutscores to determine the appropriate Stage 2 module of the examinee and name the score *target MST*.[6] As a function of the number-correct score at Stage 1, one of the three path scores will be the target MST and the other path scores will be regarded as the off-level scores.
7. Replicate Steps 1 to 7 using the large-difficulty difference MST.

All of the simulated examinees had three equated raw scores—high, middle, and low—in each of the given assembly conditions. For each path, we calculated the mean and standard deviation (SD) of the equated raw scores at each of the quadrature points in the small-difference and large-difference condition. Then we subtracted the middle path mean from the high path mean (e.g., high minus middle) and the low path mean from the middle path mean (e.g., middle minus low) for the statistical comparison. Then we calculated the conditional mean difference, the standard error of the mean difference, and $z$-statistics at each of the quadrature points.

We also calculated the difference scores by subtracting the target MST scores from the off-level scores. Therefore, negative values indicate potential underestimation caused by the off-level routing, whereas positive values indicate overestimation caused by the off-level routing. Then the mean differences were plotted by conditioning on the number-correct score at Stage 1 (routing) as a way to assess the impact of misrouting. Conditional SDs were also plotted along with the conditional mean difference to display the variability of differences across the entire score range. Furthermore, we computed the root mean squared error (RMSE)[7] at each of the number-correct score points at Stage 1.

## Results

Tables 2 and 3 present the means and SD of the equated raw scores at each of the 41 quadrature points in the small-difference and large-difference conditions, respectively. Figure 4 graphically presents the same information shown in Tables 2 and 3. In both assembly conditions, the conditional means, calculated from the 1,000 simulated examinees at each of the 41 quadrature points, were nearly identical, regardless of which module they received at Stage 2. The conditional means of the target MST that resulted from the usual MST procedure were almost identical with the means of other nontarget paths (high, middle, low in the tables). As displayed in Figure 4, the conditional mean plots of the high, middle, low, and target paths completely overlapped across the entire theta region. However, the SDs, which can be interpreted as empirical estimates of the conditional standard errors of measurement (CSEM), varied substantially depending on the choice of module at Stage 2. The SDs indicate which module led to the least random measurement error as a function of the examinees' ability. The conditional SD plots for each of the four paths varied as a function of the theta scale. The high path yielded the smallest SDs at the high theta region, whereas the low path yielded the smallest SDs at the low theta region. As expected, target MST generally led to the smallest SDs across the entire theta region. The patterns of the conditional SDs of the four paths were very similar in both assembly conditions, but their magnitudes differed as a function of the theta region and assembly conditions. At the low theta region, the high path yielded larger SDs in the large-difference condition than in the small-difference condition. At the high theta region, however, the high path yielded smaller SDs in the large-difference condition than in the small-difference condition. Similar trends appeared

**Table 2** Means and Standard Deviations of Equated Raw Scores at Each of the 41 Quadrature Points: Under the Small-Difficulty Difference Condition

| | Mean | | | | SD | | | |
|---|---|---|---|---|---|---|---|---|
| Theta | High | Middle | Low | Target MST | High | Middle | Low | Target MST |
| −3.00 | 5.81 | 5.63 | 5.45 | 5.45 | 1.99 | 1.61 | 1.38 | 1.38 |
| −2.85 | 5.96 | 6.01 | 5.69 | 5.69 | 2.09 | 1.89 | 1.55 | 1.55 |
| −2.70 | 6.49 | 6.44 | 6.25 | 6.25 | 2.32 | 2.11 | 1.78 | 1.78 |
| −2.55 | 6.76 | 6.69 | 6.57 | 6.57 | 2.52 | 2.19 | 1.93 | 1.93 |
| −2.40 | 7.32 | 7.32 | 7.17 | 7.17 | 2.66 | 2.40 | 2.03 | 2.03 |
| −2.25 | 7.98 | 7.98 | 8.02 | 8.02 | 2.84 | 2.59 | 2.20 | 2.21 |
| −2.10 | 8.62 | 8.71 | 8.73 | 8.74 | 2.98 | 2.61 | 2.19 | 2.19 |
| −1.95 | 9.38 | 9.42 | 9.60 | 9.60 | 3.02 | 2.68 | 2.21 | 2.22 |
| −1.80 | 10.24 | 10.33 | 10.58 | 10.59 | 3.08 | 2.69 | 2.15 | 2.16 |
| −1.65 | 11.33 | 11.39 | 11.62 | 11.62 | 2.91 | 2.61 | 2.09 | 2.10 |
| −1.50 | 12.14 | 12.39 | 12.57 | 12.58 | 2.94 | 2.45 | 1.99 | 2.00 |
| −1.35 | 13.13 | 13.28 | 13.52 | 13.55 | 2.80 | 2.37 | 1.88 | 1.91 |
| −1.20 | 14.41 | 14.38 | 14.53 | 14.59 | 2.57 | 2.17 | 1.83 | 1.89 |
| −1.05 | 15.42 | 15.47 | 15.53 | 15.54 | 2.35 | 1.96 | 1.83 | 1.83 |
| −0.90 | 16.17 | 16.33 | 16.41 | 16.43 | 2.34 | 1.97 | 1.78 | 1.76 |
| −0.75 | 17.30 | 17.53 | 17.50 | 17.52 | 2.20 | 1.89 | 1.76 | 1.78 |
| −0.60 | 18.36 | 18.42 | 18.52 | 18.51 | 2.15 | 1.85 | 1.78 | 1.78 |
| −0.45 | 19.50 | 19.55 | 19.58 | 19.54 | 1.98 | 1.72 | 1.76 | 1.71 |
| −0.30 | 20.48 | 20.56 | 20.67 | 20.56 | 1.88 | 1.74 | 1.77 | 1.73 |
| −0.15 | 21.45 | 21.55 | 21.58 | 21.55 | 1.88 | 1.75 | 1.80 | 1.77 |
| 0.00 | 22.54 | 22.63 | 22.64 | 22.64 | 1.84 | 1.69 | 1.80 | 1.70 |
| 0.15 | 23.60 | 23.63 | 23.66 | 23.63 | 1.78 | 1.77 | 1.86 | 1.76 |
| 0.30 | 24.52 | 24.61 | 24.64 | 24.59 | 1.73 | 1.72 | 1.96 | 1.68 |
| 0.45 | 25.67 | 25.63 | 25.73 | 25.62 | 1.80 | 1.79 | 2.00 | 1.80 |
| 0.60 | 26.56 | 26.61 | 26.64 | 26.56 | 1.68 | 1.80 | 2.00 | 1.71 |
| 0.75 | 27.56 | 27.62 | 27.70 | 27.53 | 1.69 | 1.84 | 2.09 | 1.67 |
| 0.90 | 28.55 | 28.61 | 28.66 | 28.52 | 1.75 | 1.91 | 2.08 | 1.79 |
| 1.05 | 29.59 | 29.67 | 29.67 | 29.57 | 1.77 | 1.99 | 2.38 | 1.78 |
| 1.20 | 30.47 | 30.69 | 30.70 | 30.46 | 1.68 | 2.03 | 2.41 | 1.68 |
| 1.35 | 31.45 | 31.50 | 31.61 | 31.42 | 1.87 | 2.13 | 2.54 | 1.91 |
| 1.50 | 32.36 | 32.61 | 32.65 | 32.34 | 1.82 | 2.27 | 2.67 | 1.84 |
| 1.65 | 33.33 | 33.44 | 33.55 | 33.33 | 1.97 | 2.36 | 2.76 | 1.99 |
| 1.80 | 34.16 | 34.30 | 34.39 | 34.16 | 1.91 | 2.40 | 2.76 | 1.92 |
| 1.95 | 35.14 | 35.14 | 35.28 | 35.14 | 2.02 | 2.27 | 2.72 | 2.02 |
| 2.10 | 35.88 | 35.92 | 36.00 | 35.88 | 2.04 | 2.35 | 2.59 | 2.05 |
| 2.25 | 36.60 | 36.59 | 36.71 | 36.60 | 1.95 | 2.21 | 2.47 | 1.95 |
| 2.40 | 37.17 | 37.14 | 37.19 | 37.17 | 1.78 | 2.02 | 2.29 | 1.78 |
| 2.55 | 37.69 | 37.58 | 37.65 | 37.69 | 1.69 | 1.88 | 2.13 | 1.69 |
| 2.70 | 38.08 | 37.89 | 37.96 | 38.08 | 1.50 | 1.71 | 1.95 | 1.50 |
| 2.85 | 38.36 | 38.22 | 38.20 | 38.36 | 1.35 | 1.54 | 1.76 | 1.35 |
| 3.00 | 38.68 | 38.52 | 38.51 | 38.68 | 1.13 | 1.36 | 1.60 | 1.13 |

*Note:* $n = 1{,}000$. MST = multistage test.

for the low path. At the low theta region, the low path yielded smaller SDs in the large-difference condition than in the small-difference condition. At the high theta region, however, the low path yielded larger SDs in the large-difference condition than in the small-difference condition. When the examinees received the Stage 2 module that was matched with their performance at Stage 1 (i.e., target MST), the magnitudes of the CSEMs were comparable across the entire theta range under the large-difference condition but not in the small-difference condition. In the small-difference condition, the CSEMs were generally smaller in the middle theta region, due to many overlapped items between the adjacent Stage 2 modules.

Table 4 presents the conditional mean difference of equated raw scores between two adjacent paths, along with the standard errors of the difference and $z$-statistics in the small- and large-difference assembly conditions. We

**Table 3** Means and Standard Deviations of Equated Raw Scores at Each of the 41 Quadrature Points: Under the Large-Difficulty Difference Condition

| | Mean | | | | SD | | | |
|---|---|---|---|---|---|---|---|---|
| Theta | High | Middle | Low | Target MST | High | Middle | Low | Target MST |
| −3.00 | 5.75 | 5.66 | 5.24 | 5.24 | 2.09 | 1.64 | 1.05 | 1.05 |
| −2.85 | 6.05 | 5.95 | 5.59 | 5.59 | 2.29 | 1.84 | 1.28 | 1.28 |
| −2.70 | 6.32 | 6.26 | 6.01 | 6.01 | 2.42 | 2.01 | 1.48 | 1.48 |
| −2.55 | 6.72 | 6.69 | 6.51 | 6.51 | 2.70 | 2.21 | 1.58 | 1.58 |
| −2.40 | 7.42 | 7.32 | 7.37 | 7.37 | 2.98 | 2.45 | 1.84 | 1.84 |
| −2.25 | 8.00 | 8.03 | 8.12 | 8.12 | 3.06 | 2.58 | 1.90 | 1.90 |
| −2.10 | 8.57 | 8.64 | 8.99 | 8.99 | 3.24 | 2.69 | 1.88 | 1.88 |
| −1.95 | 9.33 | 9.49 | 9.79 | 9.80 | 3.38 | 2.79 | 1.91 | 1.92 |
| −1.80 | 10.21 | 10.20 | 10.70 | 10.71 | 3.18 | 2.65 | 1.85 | 1.86 |
| −1.65 | 11.19 | 11.29 | 11.66 | 11.67 | 3.23 | 2.57 | 1.84 | 1.87 |
| −1.50 | 12.16 | 12.37 | 12.66 | 12.68 | 3.22 | 2.49 | 1.86 | 1.88 |
| −1.35 | 13.14 | 13.31 | 13.57 | 13.58 | 3.10 | 2.32 | 1.84 | 1.85 |
| −1.20 | 14.09 | 14.28 | 14.53 | 14.56 | 2.99 | 2.22 | 1.82 | 1.84 |
| −1.05 | 15.15 | 15.40 | 15.48 | 15.51 | 2.74 | 2.08 | 1.91 | 1.92 |
| −0.90 | 16.33 | 16.38 | 16.57 | 16.51 | 2.57 | 1.95 | 1.81 | 1.73 |
| −0.75 | 17.43 | 17.41 | 17.57 | 17.53 | 2.52 | 1.93 | 1.92 | 1.85 |
| −0.60 | 18.39 | 18.50 | 18.66 | 18.52 | 2.35 | 1.84 | 1.98 | 1.79 |
| −0.45 | 19.30 | 19.43 | 19.44 | 19.41 | 2.28 | 1.86 | 1.91 | 1.84 |
| −0.30 | 20.33 | 20.44 | 20.49 | 20.41 | 2.27 | 1.78 | 1.88 | 1.79 |
| −0.15 | 21.47 | 21.56 | 21.69 | 21.52 | 2.13 | 1.76 | 2.03 | 1.85 |
| 0.00 | 22.63 | 22.71 | 22.89 | 22.73 | 2.08 | 1.73 | 2.07 | 1.80 |
| 0.15 | 23.50 | 23.52 | 23.60 | 23.54 | 2.03 | 1.68 | 2.04 | 1.80 |
| 0.30 | 24.57 | 24.57 | 24.74 | 24.58 | 1.88 | 1.79 | 2.19 | 1.80 |
| 0.45 | 25.53 | 25.59 | 25.78 | 25.62 | 1.89 | 1.74 | 2.21 | 1.79 |
| 0.60 | 26.56 | 26.74 | 26.77 | 26.66 | 1.88 | 1.85 | 2.34 | 1.75 |
| 0.75 | 27.44 | 27.67 | 27.68 | 27.52 | 1.82 | 1.89 | 2.31 | 1.75 |
| 0.90 | 28.55 | 28.74 | 28.85 | 28.57 | 1.80 | 1.96 | 2.43 | 1.75 |
| 1.05 | 29.47 | 29.55 | 29.71 | 29.45 | 1.74 | 1.90 | 2.56 | 1.74 |
| 1.20 | 30.40 | 30.58 | 30.75 | 30.38 | 1.77 | 2.06 | 2.76 | 1.78 |
| 1.35 | 31.39 | 31.48 | 31.70 | 31.37 | 1.70 | 2.10 | 2.75 | 1.72 |
| 1.50 | 32.33 | 32.57 | 32.73 | 32.32 | 1.69 | 2.28 | 2.85 | 1.71 |
| 1.65 | 33.26 | 33.52 | 33.56 | 33.25 | 1.73 | 2.31 | 2.89 | 1.74 |
| 1.80 | 34.04 | 34.40 | 34.49 | 34.04 | 1.71 | 2.39 | 2.94 | 1.71 |
| 1.95 | 34.88 | 35.09 | 35.19 | 34.87 | 1.75 | 2.29 | 2.83 | 1.77 |
| 2.10 | 35.70 | 35.95 | 35.98 | 35.70 | 1.64 | 2.25 | 2.68 | 1.64 |
| 2.25 | 36.51 | 36.53 | 36.51 | 36.51 | 1.66 | 2.13 | 2.51 | 1.66 |
| 2.40 | 37.13 | 37.07 | 37.04 | 37.13 | 1.51 | 2.10 | 2.41 | 1.51 |
| 2.55 | 37.65 | 37.68 | 37.62 | 37.65 | 1.46 | 1.78 | 2.04 | 1.46 |
| 2.70 | 38.23 | 37.98 | 37.88 | 38.23 | 1.22 | 1.70 | 1.99 | 1.22 |
| 2.85 | 38.50 | 38.26 | 38.14 | 38.50 | 1.04 | 1.48 | 1.79 | 1.04 |
| 3.00 | 38.83 | 38.51 | 38.40 | 38.83 | 0.88 | 1.38 | 1.64 | 0.88 |

*Note:* $n = 1{,}000$. MST = multistage test.

focused on high versus middle at the theta region higher than −0.75, and middle versus low at the theta region lower than 0.75. Those theta regions were filled in gray. Any large score difference or variability occurring in the theta region lower than −0.5 are not likely to be of practical interest for the high and middle paths, whereas any large-differences occurring in the theta region higher than 0.5 are not likely to be of practical interest for the low and middle paths.

In Figure 5, the first plot displays the conditional mean differences between high and middle, and the second plot displays the conditional mean differences between middle and low. In each plot, the unfilled circle indicates the mean difference of equated raw scores in the small-difference condition, and the black filled triangle indicates the mean difference of those scores in the large-difference condition. The vertical line indicates the theta region where the comparison between

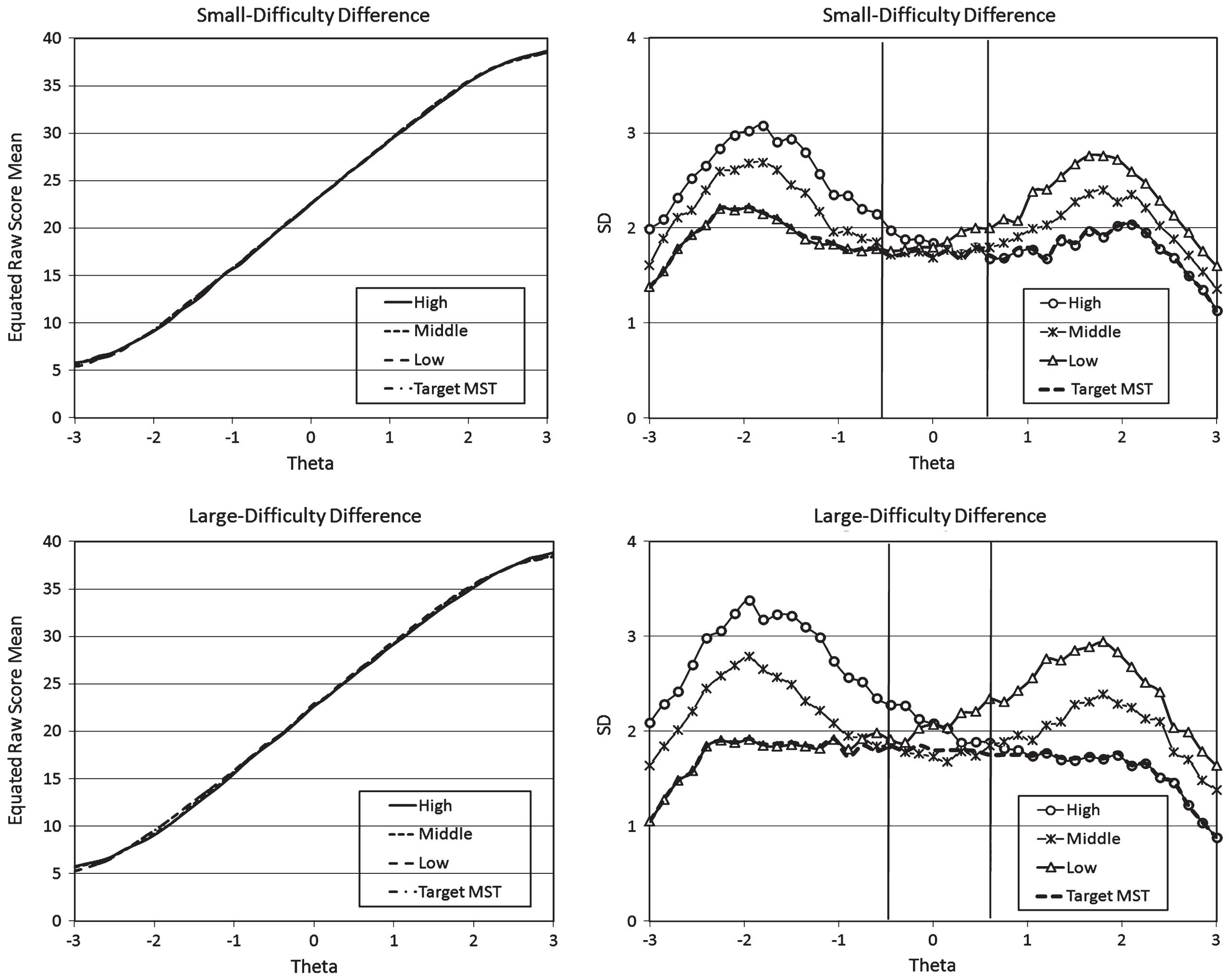

**Figure 4** Overall mean and SD (conditional standard errors of measurement) plots for the high, middle, low, and target multistage test (MST) paths under two MST assembly conditions.

the two paths would be of practical interest. In the small-difference condition, the mean differences (i.e., bias) were statistically significant (i.e., $z$-values >2.0 [or <−2.0]) in several places. The magnitude of the difference was smaller than 0.30, however, except for one place (maximum difference = 0.32, $z = 4.11$). Because biases of less than 0.5 would be of no practical interest, all the statistically significant biases were practically insignificant. The magnitudes of bias generally increased in the large-difference condition, leading to statistically significant differences in many places. The magnitude of the difference was still smaller than 0.50, however, except for one place (maximum difference = −0.51, $z = -4.97$). Although the large-difference condition led to greater bias than did the small-difference condition, the practical impact of bias was still minimal.

We compared the two sets of equated raw scores, target MST versus off-level, as another way to assess the impact of misrouting. In Figure 6, the equated raw score differences were plotted by conditioning on the number-correct raw scores at Stage 1 (routing). The conditional SD bands (difference ± 1 SD) were also plotted to display the variability of differences across the entire score range. The two plots in Figure 6 are identical except for the middle performer region. The first plot indicates the situation in which the middle performers tracked to the low (one level below) module at Stage 2 due to misrouting, whereas the second plot indicates the situation in which the middle performers tracked to the high (one level above) module at Stage 2 due to misrouting. In each plot, the black filled circle line indicates the equated score mean difference between target and off-level in the small-difference condition. The black filled triangle line indicates the same type of information in the large-difference condition. Accordingly, their mean difference ± 1 SD bands (two unfilled either circle or triangle lines) can be called the CSEM band for the difference score. The vertical lines indicate

**Table 4** Equated Raw Score Mean Differences Between Two Paths, Standard Errors of the Difference, and $z$-Statistics at Each of the 41 Quadrature Points Under the Small- and Large-Difficulty Difference Conditions

| | Small difference | | | | | | Large difference | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | High minus middle | | | Middle minus low | | | High minus middle | | | Middle minus low | | |
| Theta | Mean[a] | SE[a] | *z*[a] | Mean[b] | SE[b] | *z*[b] | Mean[a] | SE[a] | *z*[a] | Mean[b] | SE[b] | *z*[b] |
| −3.00 | 0.18 | 0.08 | 2.17 | 0.18 | 0.07 | 2.75 | 0.09 | 0.08 | 1.10 | 0.41 | 0.06 | 6.71 |
| −2.85 | −0.05 | 0.09 | −0.53 | 0.32 | 0.08 | 4.11 | 0.10 | 0.09 | 1.06 | 0.36 | 0.07 | 5.05 |
| −2.70 | 0.05 | 0.10 | 0.50 | 0.18 | 0.09 | 2.11 | 0.06 | 0.10 | 0.60 | 0.25 | 0.08 | 3.14 |
| −2.55 | 0.08 | 0.11 | 0.73 | 0.12 | 0.09 | 1.27 | 0.04 | 0.11 | 0.35 | 0.18 | 0.09 | 2.05 |
| −2.40 | 0.01 | 0.11 | 0.06 | 0.14 | 0.10 | 1.44 | 0.10 | 0.12 | 0.80 | −0.05 | 0.10 | −0.47 |
| −2.25 | 0.00 | 0.12 | −0.01 | −0.03 | 0.11 | −0.32 | −0.03 | 0.13 | −0.22 | −0.10 | 0.10 | −0.95 |
| −2.10 | −0.09 | 0.13 | −0.69 | −0.03 | 0.11 | −0.26 | −0.07 | 0.13 | −0.56 | −0.35 | 0.10 | −3.35 |
| −1.95 | −0.04 | 0.13 | −0.33 | −0.18 | 0.11 | −1.63 | −0.16 | 0.14 | −1.16 | −0.30 | 0.11 | −2.83 |
| −1.80 | −0.09 | 0.13 | −0.68 | −0.25 | 0.11 | −2.28 | 0.02 | 0.13 | 0.13 | −0.51 | 0.10 | −4.97 |
| −1.65 | −0.05 | 0.12 | −0.44 | −0.23 | 0.11 | −2.14 | −0.10 | 0.13 | −0.78 | −0.37 | 0.10 | −3.67 |
| −1.50 | −0.26 | 0.12 | −2.12 | −0.17 | 0.10 | −1.74 | −0.21 | 0.13 | −1.63 | −0.29 | 0.10 | −2.97 |
| −1.35 | −0.16 | 0.12 | −1.34 | −0.24 | 0.10 | −2.50 | −0.17 | 0.12 | −1.39 | −0.26 | 0.09 | −2.73 |
| −1.20 | 0.02 | 0.11 | 0.22 | −0.15 | 0.09 | −1.67 | −0.19 | 0.12 | −1.58 | −0.26 | 0.09 | −2.84 |
| −1.05 | −0.05 | 0.10 | −0.53 | −0.07 | 0.08 | −0.79 | −0.25 | 0.11 | −2.32 | −0.08 | 0.09 | −0.89 |
| −0.90 | −0.16 | 0.10 | −1.64 | −0.08 | 0.08 | −0.91 | −0.05 | 0.10 | −0.48 | −0.19 | 0.08 | −2.26 |
| −0.75 | −0.24 | 0.09 | −2.60 | 0.04 | 0.08 | 0.46 | 0.02 | 0.10 | 0.17 | −0.16 | 0.09 | −1.87 |
| −0.60 | −0.07 | 0.09 | −0.73 | −0.10 | 0.08 | −1.19 | −0.10 | 0.09 | −1.10 | −0.16 | 0.09 | −1.89 |
| −0.45 | −0.05 | 0.08 | −0.59 | −0.03 | 0.08 | −0.41 | −0.14 | 0.09 | −1.46 | −0.01 | 0.08 | −0.06 |
| −0.30 | −0.08 | 0.08 | −1.01 | −0.11 | 0.08 | −1.40 | −0.12 | 0.09 | −1.28 | −0.05 | 0.08 | −0.63 |
| −0.15 | −0.09 | 0.08 | −1.16 | −0.04 | 0.08 | −0.45 | −0.08 | 0.09 | −0.95 | −0.13 | 0.09 | −1.56 |
| 0.00 | −0.09 | 0.08 | −1.1 | −0.01 | 0.08 | −0.11 | −0.08 | 0.09 | −0.91 | −0.19 | 0.09 | −2.19 |
| 0.15 | −0.03 | 0.08 | −0.34 | −0.03 | 0.08 | −0.39 | −0.02 | 0.08 | −0.23 | −0.08 | 0.08 | −0.99 |
| 0.30 | −0.08 | 0.08 | −1.09 | −0.03 | 0.08 | −0.42 | 0.01 | 0.08 | 0.07 | −0.17 | 0.09 | −1.95 |
| 0.45 | 0.04 | 0.08 | 0.49 | −0.10 | 0.08 | −1.16 | −0.06 | 0.08 | −0.73 | −0.19 | 0.09 | −2.11 |
| 0.60 | −0.05 | 0.08 | −0.64 | −0.03 | 0.08 | −0.39 | −0.18 | 0.08 | −2.13 | −0.02 | 0.09 | −0.25 |
| 0.75 | −0.06 | 0.08 | −0.77 | −0.08 | 0.09 | −0.94 | −0.23 | 0.08 | −2.79 | −0.01 | 0.09 | −0.05 |
| 0.90 | −0.05 | 0.08 | −0.66 | −0.05 | 0.09 | −0.61 | −0.19 | 0.08 | −2.20 | −0.12 | 0.10 | −1.17 |
| 1.05 | −0.08 | 0.08 | −0.91 | 0.00 | 0.10 | −0.03 | −0.08 | 0.08 | −1.04 | −0.16 | 0.10 | −1.60 |
| 1.20 | −0.22 | 0.08 | −2.62 | −0.02 | 0.10 | −0.16 | −0.18 | 0.09 | −2.08 | −0.18 | 0.11 | −1.61 |
| 1.35 | −0.06 | 0.09 | −0.63 | −0.11 | 0.10 | −1.05 | −0.09 | 0.09 | −1.08 | −0.22 | 0.11 | −2.01 |
| 1.50 | −0.25 | 0.09 | −2.69 | −0.04 | 0.11 | −0.40 | −0.24 | 0.09 | −2.69 | −0.16 | 0.12 | −1.34 |
| 1.65 | −0.11 | 0.10 | −1.12 | −0.11 | 0.11 | −0.97 | −0.26 | 0.09 | −2.89 | −0.04 | 0.12 | −0.33 |
| 1.80 | −0.13 | 0.10 | −1.37 | −0.10 | 0.12 | −0.84 | −0.37 | 0.09 | −3.92 | −0.09 | 0.12 | −0.78 |
| 1.95 | 0.00 | 0.10 | 0.00 | −0.14 | 0.11 | −1.27 | −0.21 | 0.09 | −2.29 | −0.10 | 0.12 | −0.85 |
| 2.10 | −0.04 | 0.10 | −0.37 | −0.08 | 0.11 | −0.75 | −0.25 | 0.09 | −2.87 | −0.02 | 0.11 | −0.22 |
| 2.25 | 0.02 | 0.09 | 0.18 | −0.12 | 0.10 | −1.18 | −0.02 | 0.09 | −0.29 | 0.02 | 0.10 | 0.19 |
| 2.40 | 0.03 | 0.09 | 0.33 | −0.05 | 0.10 | −0.52 | 0.06 | 0.08 | 0.67 | 0.03 | 0.10 | 0.33 |
| 2.55 | 0.12 | 0.08 | 1.47 | −0.07 | 0.09 | −0.78 | −0.03 | 0.07 | −0.42 | 0.05 | 0.09 | 0.62 |
| 2.70 | 0.19 | 0.07 | 2.64 | −0.07 | 0.08 | −0.89 | 0.25 | 0.07 | 3.78 | 0.10 | 0.08 | 1.15 |
| 2.85 | 0.14 | 0.06 | 2.17 | 0.02 | 0.07 | 0.25 | 0.25 | 0.06 | 4.31 | 0.12 | 0.07 | 1.57 |
| 3.00 | 0.16 | 0.06 | 2.92 | 0.01 | 0.07 | 0.15 | 0.32 | 0.05 | 6.19 | 0.11 | 0.07 | 1.67 |

*Note:* $n = 1{,}000$.

[a] Data in this column ranging from a theta of 0.60 to a theta of 3.00 (shown by shading) were the focus.

[b] Data in this column ranging from a theta of −3.00 to a theta of 0.60 (shown by shading) were the focus.

two cutscores that allocated the examinees into one of the three paths: low, middle, or high. As shown in the previous comparisons, the mean differences were almost negligible across the Stage 1 score scale, indicating almost no bias. The CSEM bands of the difference scores were slightly wider than the ones of the target MST path (see Tables 2 and 3). The score variability caused by misrouting (random error plus some levels of systematic error) was slightly greater than the score variability expected by random measurement error. The same trend appeared for both assembly conditions. As shown in Figure 7, the RMSE derived from the two assembly conditions were almost indistinguishable, as in the bias and CSEM bands.

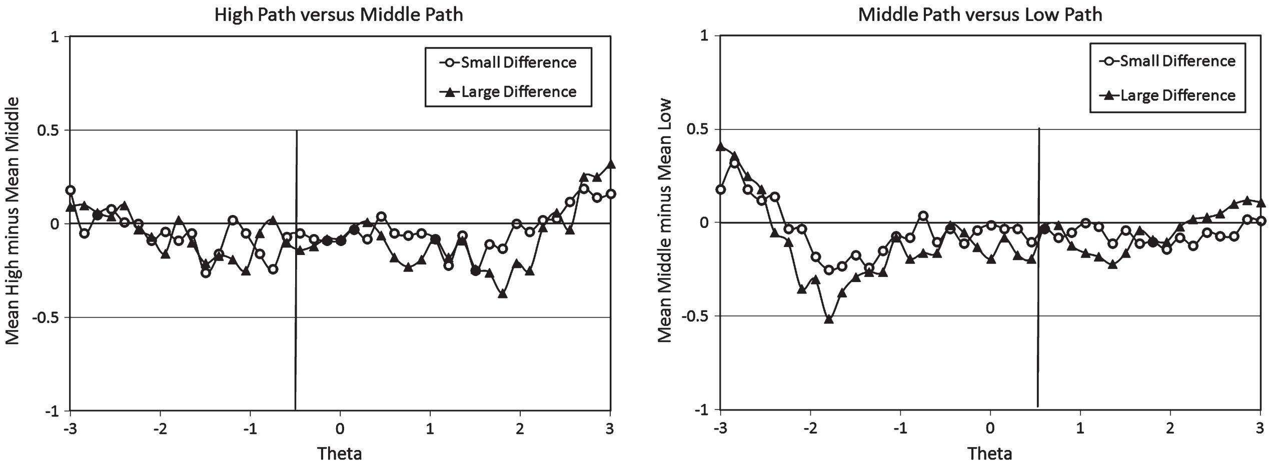

**Figure 5** Any two adjacent paths' mean difference plots under two multistage test assembly conditions.

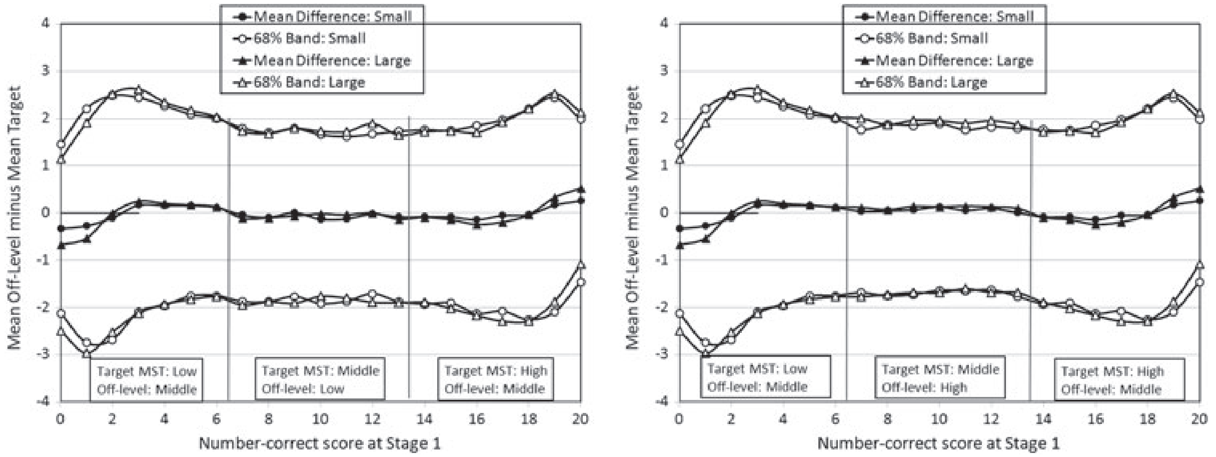

**Figure 6** Mean and 68% SD band of the difference score between target multistage test (MST) and off-level path under two MST assembly conditions.

## Conclusions

Under MST, the between-stage adaptation to match examinees' ability level to a particular level of modules is crucial in estimating the examinees' ability and in reporting accurate scores to them. Therefore, it is useful to know the extent to which the examinees' scores would change if they received a module that did not match their ability. The purpose of this study was to explore this issue using simulated data sets.

In our simulation, the impact of misrouting was minimal. Although bias increased when the three modules at Stage 2 were significantly distinct in difficulty, the mean differences between any adjacent paths were generally smaller than 0.5 across all the score regions for both assembly conditions. On average, the differences between the target and off-level scores were also trivial for practical purposes. The CSEM bands of the difference scores between any adjacent paths were only slightly wider than the CSEM bands of the target MST path. Misrouting might be beneficial to some examinees but harmful to others. More importantly, however, there were no systematic trends that raised any concerns about the scores' fairness.

Any large score differences or variability occurring at either the low or high region of the scale are not likely to be of practical interest, because differences in those regions will rarely result in inaccurate assignments in actual testing situations. As the RMSE plots indicate, total error associated with misrouting was smaller near the two cutscore regions than in the high and low regions of the scale. The variability (1.8 – 1.9) observed near the two cutscore regions was comparable with the CSEM (random variability; 1.7 – 1.8) of the target MST path.

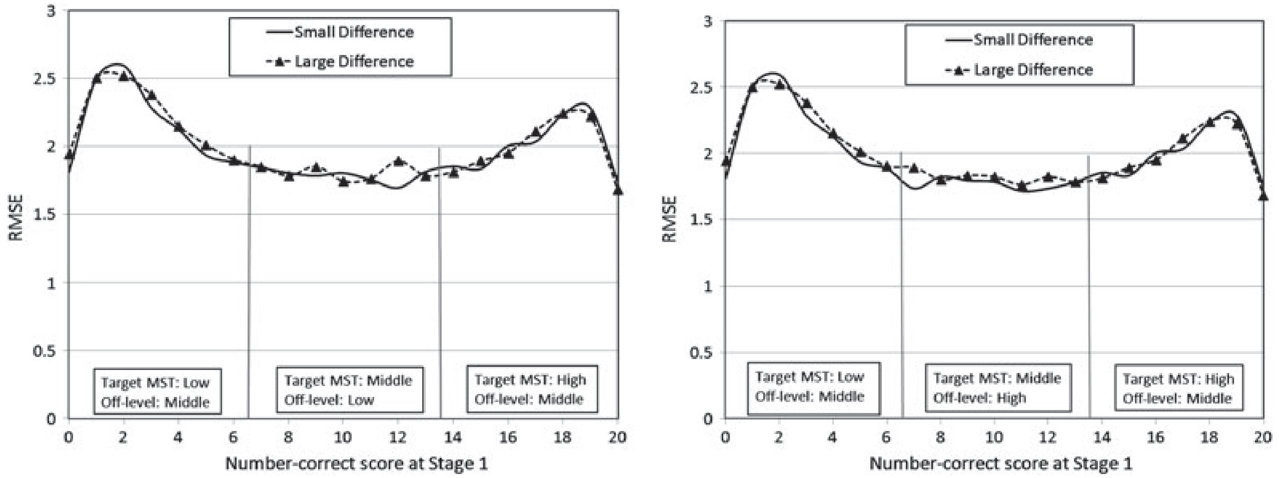

**Figure 7** Root mean squared error (RMSE) of the difference score between target multistage test (MST) and off-level path under two MST assembly conditions.

In general, examinees' final scores would change depending on the choice of routing. On the basis of the findings from the current simulation, however, it is concluded that the magnitude of score change caused by any systematic routing error will be comparable with the score changes caused by random measurement error, even for borderline examinees whose routing scores fall near the cutscore regions. Estimates of examinees' ability may be robust enough to shield them from the effects of misrouting, particularly in the number-correct scoring two-stage MST design used in this study, unless MST modules include many misfitting and unreliable items.

In contrast to our expectations, the impact of misrouting was not clearly associated with the statistical characteristics of Stage 2 modules. Bias in the large-difference condition was larger than that in the small-difference condition. However, the increment in bias was minor compared to the magnitude of difference in difficulty imposed on both high and low modules in the large-difference condition (1 SD unit above or below). Although total misrouting error was nearly identical in both assembly conditions (shown in Figure 7), the comparison between the two MST assembly conditions shows a clear benefit from the use of distinct Stage 2 modules under a two-stage MST. The overall test score reliability and standard errors of measurement were almost identical for both assembly conditions. The CSEM patterns differed slightly, however, between conditions. In the small-difference condition, the CSEM of the target MST was smaller near the theta region from −1.0 to +1.0, because the routing module at Stage 1 included many medium-difficulty items rather than either hard or easy items, and the second-stage high and low modules overlapped substantially with the middle module in difficulty. In the large-difference condition, however, the CSEM of the target MST was fairly similar in size across the entire score scale, because many high module items targeted strong performers, whereas many low module items targeted weak performers. Maintenance of similar levels of measurement precision across the entire score scale is promising, particularly for assessments in which all score points are equally important. In addition, the CSEM of the target MST was slightly smaller in the large-difference condition than in the small-difference condition.

The choice of MST design configurations and psychometric characteristics of MST assembly are influenced by various factors, such as test score uses (certification or noncertification), item security, item pool capacity, administration environments, and so forth. When the number of items for a routing module is small (e.g., fewer than 20), item pool capacity is limited, and few adaptations are available in the MST design, it may be safe to assemble modules that overlap in difficulty. This practice, however, may offset some degree of measurement precision at a particular score region. As shown in this study, the use of modules that are clearly distinct in difficulty at subsequent stages will have psychometric benefits. Note that such assembly design may challenge test developers because, in most testing programs, writing well-performing difficult questions apparently is harder than writing easy or medium-difficulty items.

Concerns have been raised about the potential score variability caused by routing error, particularly for borderline examinees whose scores are near the routing cutscores. In reality, however, those borderline examinees' scores would not be heavily influenced by the choice of subsequent stage module. Perhaps potential differences in reported scores caused by routing itself are minimal. More dramatic changes would appear if clearly strong or clearly weak performers received a module that was not best matched with their actual ability levels. In reality, however, such cases would rarely occur. The present simulation supports these arguments, as does a previous simulation study by Armstrong (2002). Perhaps, in practice, the impact of routing error will not be substantial, as long as practitioners make an effort to ensure the quality of MST (e.g., use of secure and highly discriminating items) and the proper execution of item calibration and linking.

In this simulation, we examined a two-stage MST design using the number-correct scoring method, because an international testing program recently adopted the same conditions for operational use. Further simulation studies would be useful for clarifying the importance of routing by manipulating various measurement conditions, particularly at the routing stage (e.g., optimal number of items in a routing module). Comparisons of various scoring methods (i.e., proficiency estimation) would be interesting in helping to determine which scoring method performs well despite misrouting. In the current simulation, we considered only the second-stage modules' difference in difficulty, not in discrimination. It would be worthwhile to consider item discrimination factors as well to see if using more difficult (or easier) but less discriminating items would confirm the current findings. Particularly, comparing number-correct scoring with item-pattern scoring would be interesting in a situation in which misrouting takes place but poor-quality items are included in subsequent stage modules.

## Notes

[1] Recently, an international testing program adopted this MST design for operational use.

[2] The 2PL model is used operationally for some large-scale testing programs (e.g., the *GRE®* and *TOEFL®* testing programs). To make our simulation as realistic as possible, we examined item parameter statistics derived from over 100 operational two-stage MST forms before specifying the descriptive statistics (e.g., mean, SD, min, and max) of item difficulty and item discrimination parameters for each of the four modules.

[3] We used a constant discrimination parameter ($a = .80$) for all items.

[4] The *defined population intervals* method can be used to implement a policy that specifies the relative proportions of examinees in the population expected to follow each of the available paths through the panel. Under the cutscores and ability distribution conditions employed in this simulation, the ability scores associated with the 30th and 70th percentiles of the cumulative distribution of theta would be approximately -.53 and + .53. This can be verified easily from a standard table of values for the unit normal distribution. The approximate number-correct routing scores at Stage 1 could then be determined through the TCC under summed scoring.

[5] Location of routing points may be more appropriate under the small-difference condition than under the large-difference condition. However, any difference caused by use of slight difference cutscores would be insignificant under summed scoring.

[6] We used a Stage 1 raw score of 6 to differentiate the low performers from the middle performers, and a Stage 1 raw score of 14 to differentiate the middle performers from the high performers.

[7] $\mathrm{RMSE}_{\mathrm{Theta}} = \sqrt{\mathrm{Mean}^2_{\mathrm{Theta}} + \mathrm{SD}^2_{\mathrm{Theta}}}$

## References

Armstrong, R. D. (2002). *Routing rules for multiple-form structures* (Computerized Testing Report No. 02-08.) Newtown, PA: Law School Admission Council.

Educational Testing Service. (2011). *GRE information and registration bulletin*. Princeton, NJ: Author. Retrieved from http://www.ets.org/s/gre/pdf/gre_info_reg_bulletin.pdf

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice*, *26*, 44–52.

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

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

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

Luecht, R. M., Nungester, R. J., & Hadidi, A. (1996, April). *Heuristic-based CAT: Balancing item information, content, and exposure.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Luecht, R. M., & Sireci, S. G. (2011). *A review of models for computer-based testing* (Research Report No. 2011-12). New York, NY: The College Board.

Melican, G. J., Breithaupt, K., & Zhang, Y. (2010). Designing and implementing a multistage adaptive test: The uniform CPA examination. In W. J. van der Linden & C. E. W. Glas (Eds.), *Elements of adaptive testing* (pp. 167–189). New York, NY: Springer.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York, NY: Springer.

Wang, X., Fluegge, L., & Luecht, R. (2012, April). *A large-scale comparative study of the accuracy and efficiency of ca-MST.* Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, BC, Canada.

Wise, S. L. (1996, April). *A critical analysis of the arguments for and against item review in computerized adaptive testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

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

**Action Editor:** Marna Golub-Smith

**Reviewers:** Xueli Xu and Tsung-Han Ho

ETS, the ETS logo, GRE, LISTENING. LEARNING. LEADING., and TOEFL are registered trademarks of Educational Testing Service (ETS). All other trademarks are property of their respective owners.

Find other ETS-published reports by searching the ETS ReSEARCHER database at http://search.ets.org/researcher/