JEM JOURNAL of EDUCATIONAL MEASUREMENT
NCME national council on measurement in education
Journal of Educational Measurement
Spring 2019, Vol. 56, No. 1, pp. 192–213

Routing Strategies and Optimizing Design for Multistage Testing in International Large-Scale Assessments

Dubravka Svetina
Indiana University
Yuan-Ling Liaw
University of Oslo
Leslie Rutkowski and David Rutkowski
Indiana University

This study investigates the effect of several design and administration choices on item exposure and person/item parameter recovery under a multistage test (MST) design. In a simulation study, we examine whether number-correct (NC) or item response theory (IRT) methods are differentially effective at routing students to the correct next stage(s) and whether routing choices (optimal versus suboptimal routing) have an impact on achievement precision. Additionally, we examine the impact of testlet length on both person and item recovery. Overall, our results suggest that no single approach works best across the studied conditions. With respect to the mean person parameter recovery, IRT scoring (via either Fisher information or preliminary EAP estimates) outperformed classical NC methods, although differences in bias and root mean squared error were generally small. Item exposure rates were found to be more evenly distributed when suboptimal routing methods were used, and item recovery (both difficulty and discrimination) was most precisely observed for items with moderate difficulties. Based on the results of the simulation study, we draw conclusions and discuss implications for practice in the context of international large-scale assessments that recently introduced adaptive assessment in the form of MST. Future research directions are also discussed.

Recognizing the advantages of computer-based assessment (CBA; Jodoin, Zenisky, & Hambleton, 2006; Yan, Lewis, & von Davier, 2014b, p. 4), the Organization for Economic Cooperation and Development (OECD) implemented an optional CBA in the 2012 cycle of their flagship study, the Programme for International Student Assessment (PISA; OECD, 2014). In particular, the OECD cited CBAs as a way to "make the assessment process more efficient and narrow the time lag between collecting the data and making results available to feed into educational improvement" (OECD, 2010, p. 4). Given that dozens of highly heterogeneous educational systems take part in PISA and other international large-scale assessments (ILSAs), a computerized platform offers a further advantage: the potential to include an adaptive element.

Beginning in 2011, the OECD's Program for the International Assessment of Adult Competencies (PIAAC; Kirsch & Lennon, 2017) implemented the first adaptive ILSA in the form of a multistage test (MST). Following PIAAC's lead, the 2018 cycle of PISA also features an MST component (Educational Testing

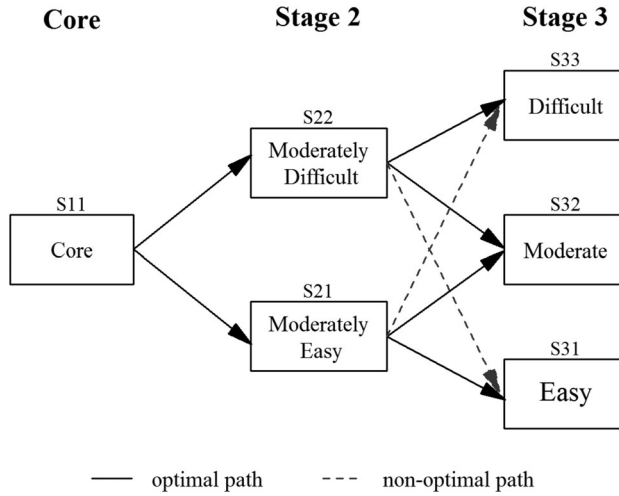

Figure 1. An example of a 1-2-3 three-stage MST design used in the study.

Service, 2016). Two key reasons for this innovation include an improved assessment experience for students and more accurate and valid population estimates for many participating systems (OECD, 2010, p. 4). In spite of the advantages of an MST design over a more traditional linear fixed-length test, implementing this sort of approach presents unique challenges in an ILSA context. For example, PIAAC faced administration challenges because the adaptive algorithms needed to be sufficiently flexible to account for population diversity while ensuring even exposure of test booklets across populations (Chen, Yamamoto, & von Davier, 2014). Further issues, discussed subsequently, involve test stage length and scoring. In the current study, we are interested in even item exposure within a population, with an interest in examining these issues in multiple populations in the future. It is in this context that the current article is written.

Background

MST is a design that allows for limited adaptation of a test's difficulty to the proficiency of the examinee. It is limited in that, by comparison to a fully computerized adaptive test (CAT), adaptation happens not at the item level but instead at the *module* or testlet level. Here, module is defined as a group of items that always appear together in a block or cluster. One example of an MST is presented in Figure 1. Here, a three-stage MST structure, referred to collectively as a *panel*, shows the various paths that an examinee may take.[1] Examinees begin in Stage 1 with a core or routing testlet (S11). Depending on the score in S11, an easier or more difficult Stage 2 testlet is selected, either S21 or S22. Again, based on performance in the previous stage(s), the examinee is routed into a Stage 3 testlet. In total, each examinee receives three blocks of nonoverlapping items. Although the above description is limited to a single panel, an MST can be composed of many panels. Further, the described design is just one of many possible in the MST framework.[2]

Scholars have pointed to several advantages of an MST design when compared to other testing formats. For example, when compared to linear fixed-length tests, MST showed greater testing efficiency (Jodoin et al., 2006; H. Kim & Plake, 1993) and increased accuracy in ability estimates (and classification). Moreover, while not as efficient as a CAT, MST may still be a desirable operational choice (Melican, Breithaupt, & Zhang, 2009). As Luo and Kim (2018) suggested, MST provides several practical advantages over CAT (Melican et al., 2009), including the *a priori* knowledge of psychometric and content properties of all possible test forms, as MST is constructed prior to administration; a more efficient approach to deal with complex test constraints and thus reduce computing power; and the flexibility for the test taker to review and revise responses in the same stage of the assessment.

Although MST theory and practice is fairly well-established for tests used to make individual decisions, the novelty of this method in an ILSA setting implies several open research questions. As noted above, controlling item exposure in PIAAC posed operational challenges. Although controlling item *over*-exposure in many adaptive tests is desirable for test security (Stocking & Lewis, 1998), even item exposure in an ILSA setting is necessary to ensure that item parameter estimates are not biased. This is because item parameters are estimated subsequent to data collection and exposure to limited subsets of the tested populations, which, in PISA, number more than 80, raises concerns about the stability of parameter estimates. In response, Chen and colleagues developed a probabilistic routing procedure that relied on examinee background information known *a priori*, including their native language and education levels, as well as performance on previous testlets (Chen et al., 2014). This approach ensured that, regardless of background and prior performance, examinees had a nonzero probability of being routed into any of the available subsequent testlets, offering some safeguards against item over- or underexposure. In contrast to PIAAC, most other ILSAs, including PISA and the Trends in International Mathematics and Science Study (TIMSS) collect background information on students *ex post facto*, eliminating the possibility of using student background as part of a routing scheme. Especially in a setting where educational systems differ meaningfully across the proficiency spectrum, an open question regards how best to ensure even testlet exposure. For example, given a historic proficiency mean and standard deviation of 500 and 100 (Mullis, Martin, Foy, & Hooper, 2016), respectively, the difference between the highest and lowest 2015 grade eight TIMSS mathematics performers was just over 2.5 standard deviations (Singapore, $\bar{x} = 621$, Saudi Arabia, $\bar{x} = 368$). Such wide differences suggest that in an MST setting that uses a routing scheme based only on performance will underexpose difficult items and overexpose easy items in low-performing systems and vice versa in high-performing systems.

A second issue involves the method used to score testlets for routing purposes. Routing to the next optimal testlet is often, although not always, based on one of two general approaches: (a) a number-correct (NC) scoring approach (i.e., *classical*) using some predetermined observed score cutoff value (and/or associated percentage of passing/not passing examinees); or (b) an item-pattern (item response theory [IRT]) approach such as using maximum information or an interim ability estimate[3] (Hendrickson, 2007; Yan, von Davier, & Lewis, 2014). As S. Kim, Moses, and Yoo (2015) suggested, the practical benefits of NC scoring, including being easier for

test takers to understand (e.g., Armstrong, 2002), versus the psychometric benefits of item-pattern scoring, such as more precise estimation, have been debated in the literature. For example, Luecht and Nungester (1998) empirically showed that NC scoring can be sufficiently accurate to select testlets. Similarly, although Robin, Manfred, and Liang (2014) found that NC yielded loss in measurement accuracy in their preliminary study (when compared to maximum likelihood estimates), it presented small observed losses (across the most of the latent trait continuum of 5% or less, and 10% of loss in the upper and lower ends of the continuum). Practical support for an NC method was also reported by Weissman, Belov, and Armstrong (2007), who found that, although information-based methods yielded higher overall classification rates over NC, it came at the expense of item (over)exposure, particularly in later MST stages. Still others found that shrinkage is more pronounced under NC scoring due to its lower precision than item-pattern scoring such as expected *a posteriori* (EAP; e.g., Kolen & Tong, 2010). The above literature was situated in the context of high-stakes tests used for decision making. To our knowledge, there is little ILSA-based research on this question, where inferences are limited to the population and subpopulation level (Mislevy, Johnson, & Muraki, 1992). A particular focus is on the way in which performance-based and probabilistic routing decisions might interact. We pursue this question here.

A third issue considered in the current article is one aspect of test assembly. To that end, Yan, Lewis, and von Davier (2014a) examined dimensions of assembly in the context of a small sample size study with few items to study the performance of regression-tree-based scoring. In the same volume, Zheng, Wang, Culbertson, and Chang (2014) review several assembly methods, including a number of automated approaches. In both studies, the findings are not well connected to the ILSA setting, where, again, item parameters are only known subsequent to data collection (although preliminary estimates are available from field trials) and sample sizes are large (hundreds of thousands of students are available for international item calibration). As such, we consider one aspect of testlet and panel assembly here: that of testlet length. In other words, we consider whether having testlets of equal or unbalanced lengths demonstrates advantages.

As ILSAs move into MST, it is timely that these design and administration choices are more systematically considered. It is in this context that we situate our article. Specifically, we examine the effect of several design and administration choices on item exposure and person-parameter recovery. In order to address the main goals of the study, we utilize a Monte Carlo simulation approach. In the next section, we describe the methods used in the current study with an emphasis on the implemented study design, rationale for selected design choices, and the outcome variables.

Methods

We utilize a Monte Carlo simulation study to address the research questions in our study via the *R* (R Development Core Team, 2018) package *mstR* 1.2 (Magis, Yan, & von Davier, 2018)[4] for MST simulation and analyses and the *mirt* package (Chalmers, 2012) for item parameter calibration. In the study, several aspects were treated as fixed: the sample size ($n = 4{,}000$ simulees), for which abilities were

generated from a normal distribution $N(0, 1)$; number of items per design was fixed at 36; and the MST was set as a *1-2-3* three-stage design (as presented in Figure 1).[5] One hundred replications were performed within each condition. Our sample size is generally reflective of operational ILSA settings, where sample sizes range from 3,000 to 8,000 or more students per tested population.

Several manipulated factors were included in the study: (a) the number of items per testlet, (b) routing method to the next testlet, and (c) routing probabilities. Next, we elaborate on the study design, including the manipulated factors (and respective levels), provide a rationale for the design choices as driven by the ILSA context, and outline the data generation and analysis plan, including outcome variables that align with our research goals.

Manipulated Factors

Number of items per testlet. We manipulated four sizes that any one testlet could assume. Recall that our design involved a 1-2-3 MST form. We were particularly interested in examining different lengths of testlets because of our interest in balanced item exposures and parameter recovery. Across all stages, either 6, 10, 12, or 20 items per testlet were selected in any condition, while maintaining a fixed number of 36 items for all simulees. This resulted in four designs, with balanced and imbalanced testlet lengths:

(1) Design 1 (*equal*) had 12 items in each of the three testlets (*EQ*; 12-12-12 items),
(2) Design 2 (*short-to-long*) had six items in the Core testlet, followed by 10 and 20 items in subsequent testlets (*S-L;* 6-10-20 items),
(3) Design 3 (*long-to-short*) had 20 items in the Core testlet, followed by 10 and 6 items in subsequent testlets (*L-S*; 20-10-6 items), and
(4) Design 4 (*short-long-short*) had six items in the Core testlet, followed by 20 and 10 items in subsequent testlets (*S-L-S*; 6-20-10 items).

Routing method. In the current study, we utilized five different routing methods in order to select the optimal next testlet:

(1) random selection, which meant that examinees were distributed to the next testlet with equal probability regardless of their provisional performance,
(2) NC cumulative score (i.e., using a cutoff value based on total score),
(3) NC last testlet score (i.e., using a cut off value based on the last administered testlet only, rather than performance on all previously administered testlets),
(4) IRT EAP (i.e., IRT EAP ability estimate), and
(5) IRT information (i.e., IRT maximum Fisher information function).

Cutoff values for the routing methods (2) and (3) for two adaptations are listed in Table 1. For example, under the *EQ* design, there are 12 items in the Core testlet, so a simulee could earn anywhere between 0 points (no correct answer) to 12 points (all correct answers). If a simulee scored 6 or below on the Core testlet, they would be routed to the lower testlet in Stage 2 (labeled as S21 in Figure 1), otherwise they would be routed to the higher testlet in Stage 2 (labeled as S22 in Figure 1).

Table 1
Ranges of Values for Classical Approaches in Selecting the Next Optimal Testlet

Design	Stage 1	Stage 2	Stage 3	1st Routing Range Values	2nd Routing Range Values
1. EQ	12	12	12	(0,6) (7,12)	(0,7) (8,16) (17,24)
2. S-L	6	10	20	(0,3) (4,6)	(0,5) (6,11) (12,16)
3. L-S	20	10	6	(0,10) (11,20)	(0,9) (11,20) (21,30)
4. S-L-S	6	20	10	(0,3) (4,6)	(0,8) (9,17) (18,26)

Note. Design 1: EQ (equal) had 12 items in each of the three testlets (12-12-12 items), Design 2: S-L (short-to-long) had six items in the Core testlet, followed by 10 and 20 items in subsequent testlets (6-10-20 items), Design 3: L-S (long-to-short) had 20 items in the Core testlet, followed by 10 and 6 items in subsequent testlets (20-10-6 items), and Design 4: S-L-S (short-long-short) had six items in the Core testlet, followed by 20 and 10 items in subsequent testlets (6-20-10 items). Routing range values represent cutoff values used in selecting the next module. For example, under EQ design, the first routing implies that if a simulee obtains a total score in Stage 1 (Core) module of 6 or below (out of 12), they would be routed to S21 module in Stage 2. If they obtained a score 7 or above (out of 12), they would be routed to S22 in Stage 2.

Routing probabilities. To investigate the effect on item exposure and parameter recovery, we consider a probabilistic routing mechanism, where, regardless of performance, a student can be routed to either an optimal or suboptimal testlet, with some nonzero probability (e.g., dashed lines in Figure 1). In other words, a student who answers all items correctly and *should* be routed to a more difficult panel in the stage would have a predetermined probability of being routed into an easier stage, ensuring that some portion of high achievers would be exposed to easy items.

Levels of this manipulated factor included optimal testlet routing probabilities of (a) 1.00, (b) .80, or (c) .70. When probability equals 1.00, all simulees are optimally routed based on performance only. In the other two conditions, all simulees face a .20 or .30 probability of routing to a suboptimal testlet regardless of performance. For example, optimal routing from Stage 1 to Stage 2 has a probability of .80 or .70, respectively, while suboptimal routing happens with a probability of .20 or .30, respectively. An identical procedure is implemented for routing from Stage 2 to Stage 3. A distinction in the second routing procedure is that when suboptimal routing is selected, either of two possible testlets is selected with equal probability.

When the routing probability was 1.00 for an optimal testlet, there existed four possible paths: Path 1: Core – Moderately Easy – Easy; Path 2: Core – Moderately Easy– Moderate; Path 3: Core – Moderately Difficult – Moderate; Path 4: Core – Moderately Difficult – Difficult. Introducing a probabilistic element into routing produced an additional two paths to which simulees could be routed. That is, when the routing probability was less than 1.00, the two additional paths (marked as dashed lines in Figure 1) were: Path 5: Core – Moderately Easy– Difficult, and Path 6: Core – Moderately Difficult – Easy.

Data Generation and Analysis

In simulating data, we selected item parameters with specific ranges for each testlet. The Core testlet at Stage 1 was of medium difficulty where item difficulty

(location) parameters were selected randomly between –1 and 1 logits. At Stage 2, item difficulty parameters were selected randomly between .5 and 1.5 for the moderately high difficulty testlet and between –1.5 and –.5 for moderately low difficulty testlet. At Stage 3, the item difficulty parameters for the highest, medium, and lowest level of difficulty testlets were randomly selected from 1 to 2, from –1 to 1, and from –2 to –1 logits, respectively. Item discrimination for items were randomly sampled between .5 and 1.5 logits for all items, balancing similar levels of discriminations across all testlets. Table 2 displays mean and variances for item difficulty and discrimination parameters used in data generation. These item parameters are generally reflective of empirical parameters observed in PISA (OECD, 2017) and were chosen to reflect differences between the more difficult and easier testlets. All item responses were generated under the dichotomous two-parameter logistic (2PL) IRT model using above-discussed item parameters and randomly selected person parameters from a standard normal distribution.

The optimal panels (a)–(d) in Figure 2 show the results for the test information function constructed from four different paths for each test length design. Under the same path, different test length designs yield parallel test forms. The peaks of the information functions for the four paths are located at approximately –1.14, –0.69, 0.28, and 0.89, respectively, on standard normal proficiency scale. And the maximum values of the information functions are 12.39, 12.43, 11.79, and 12.10. On the other hand, panels (e) and (f) display the results of two nonoptimal paths with information functions that are more irregular. In terms of Path 5, the peaks for each test length design are located at –0.60, 1.27, –0.62, and –0.97, respectively, on a theta scale and the maximum values of path information function ranged from 9.16 to 10.91. As for Path 6, the peaks are located at 0.02, –1.36, 0.45, and 0.75 and the maximum values of path information function ranged from 9.44 to 10.88. Taken all together, Figure 2 shows a contrasting breath and height of information functions between the optimal (panels a through d) and nonoptimal (panels e and f) paths. Namely, the height of information functions was higher for optimal paths than nonoptimal paths, while the breadth of the information function was larger for nonoptimal paths than it was in optimal path counterparts.

Evaluation Criteria

As stated above, the twofold purpose of the study was to examine person/item parameter estimates and item exposure. With respect to ability estimation, we evaluated performance by computing the average bias and root mean square error (RMSE) for the parameter estimates across 100 replications within each condition, as well as correlations between the estimated and true values of theta. Item parameter bias was defined as the average difference between the estimated and true values of the parameters across 36 items, while the RMSE was obtained by taking the square root of the mean of squared deviations of estimated parameter values about their true values. With respect to item exposures, we computed item exposure rates as the ratio between the number of times an item was administered and total number of simulees. We also examined item parameter recovery to examine whether items' calibration was affected by items' exposure rates.

Table 2
Means and SDs of Item Difficulty (Panel a) and Discrimination (Panel b) Parameters in Each Testlet and for Each MST Design

	Stage 1		Stage 2				Stage 3					
	Testlet 1		Testlet 21		Testlet 22		Testlet 31		Testlet 32		Testlet 33	
Design	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
						Panel (a) Difficulty Parameters						
1. EQ	−0.117	0.545	−1.192	0.265	0.712	0.154	−1.579	0.312	−0.477	0.396	1.338	0.300
2. S-L	0.022	0.789	−1.128	0.252	0.746	0.273	−1.621	0.302	−0.243	0.604	1.379	0.302
3. L-S	−0.005	0.612	−1.223	0.240	0.746	0.273	−1.707	0.270	−0.114	0.819	1.272	0.232
4. S-L-S	0.022	0.789	−1.121	0.302	0.879	0.302	−1.700	0.251	−0.209	0.619	1.519	0.333
						Panel (b) Discrimination Parameters						
1. EQ	1.200	0.424	1.200	0.318	1.200	0.534	1.200	0.461	1.200	0.389	1.200	0.433
2. S-L	1.200	0.415	1.200	0.399	1.200	0.547	1.200	0.415	1.200	0.449	1.200	0.416
3. L-S	1.200	0.462	1.427	0.369	1.200	0.547	1.200	0.499	1.200	0.314	1.205	0.371
4. S-L-S	1.200	0.415	1.200	0.387	1.200	0.501	1.200	0.350	1.200	0.426	1.200	0.492

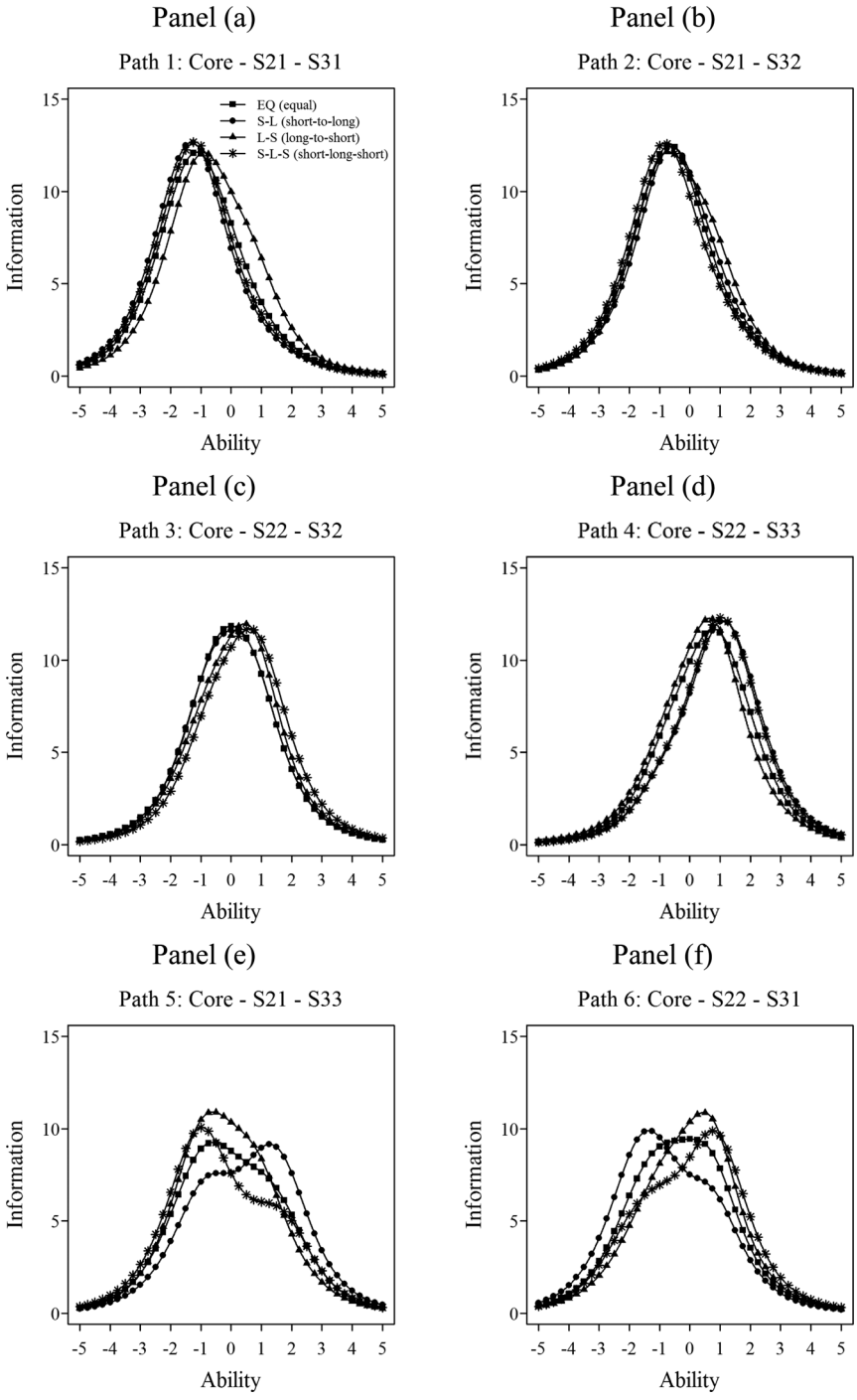

Figure 2. Test information function by optimal path (panels a–d) and nonoptimal path (panels e–f).

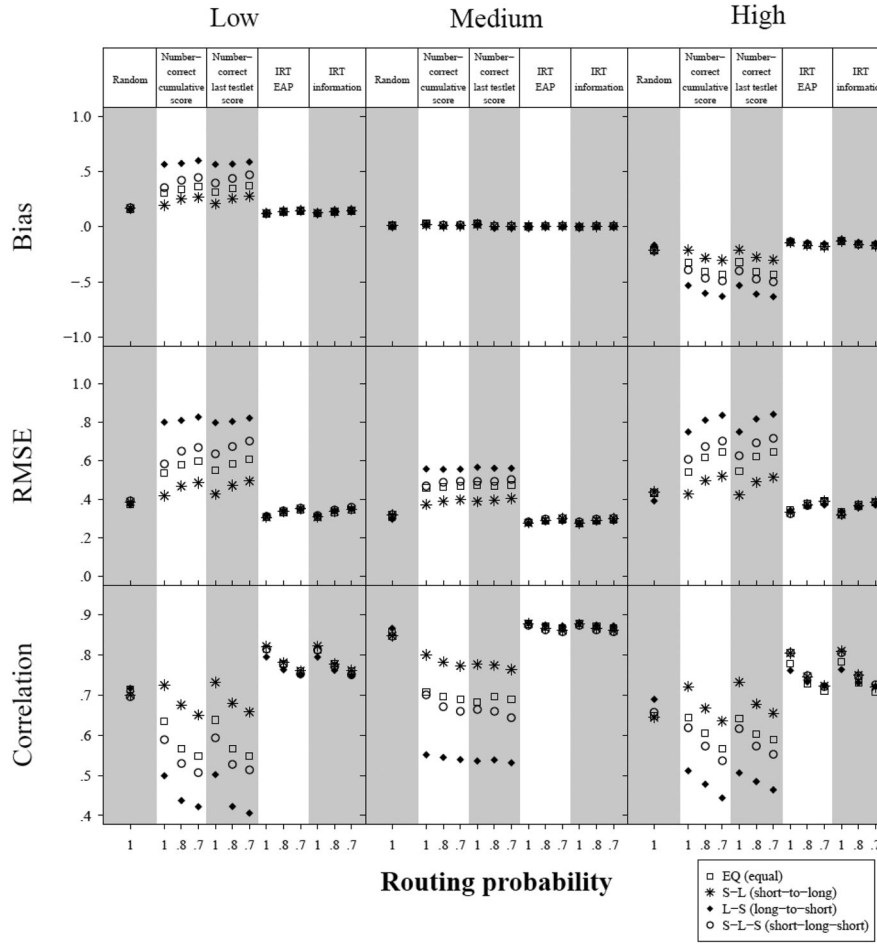

Figure 3. Person parameter recovery rates of bias, RMSE, and correlations for low, medium, and high performers.

Results

We report results in two sections, with Section I presenting results for proficiency estimation recovery, while Section II discusses results related to item exposure and recovery. For space reasons, we present selected results to highlight the main findings; however, all tabulated and expanded graphical results can be found at https://figshare.com/s/a9345b3c71a8b973630f.

Section I. Person Ability Parameter Recovery

In Figure 3, results related to the recovery of person parameter are presented. In the figure, rows represent three different outcome variables: bias, RMSE, and correlations between the true and estimated thetas. Columns represent different groups of simulees. Results reported are (albeit arbitrarily) divided into three groups: low

(estimated theta values are below –1), moderate (estimated theta values are between –1 and +1), and high (estimated theta values are greater than +1). Within each figure, results for four designs are represented as various markers (e.g., squares represent the EQ design, which had 12 items in each of the three testlets).[6] Further, in each of the graphs, the *x*-axis represents the five different methods used as the basis for testlet selection (e.g., random, NC cumulative score, NC last testlet score, IRT EAP, and IRT information), while the *y*-axis represents the three outcomes, respectively. Finally, the *x*-axis ticks on the bottom indicate the various studied probabilities for routing.

Focusing on bias, it was noted that the IRT-based methods (EAP and information) and random routing were most successful in recovering the theta estimates across the three group levels for all designs and probability routing conditions. Across the conditions, bias ranged from –0.001 to 0.001 for the IRT-based methods. Unlike the IRT methods, however, slightly larger bias was found under the NC methods, although the bias was inverse for the low- versus the high-performing group. Specifically, bias was positive for low performers, while negative bias of similar magnitudes was found for high performers. Two further patterns were noted. First, for the NC methods, differences in designs were noted across the probability levels. Namely, the smallest levels of bias were found under *S-L* design while the largest bias was found under the *L-S* design. Second, for the low/high performing group, as the routing probabilities decreased from 1.0 to .8 and .7, bias slightly increased/decreased for all four designs. We noted that for the medium group (in the –1 to +1 theta range) recovery was near perfect across all methods and routing probabilities.

The second row of Figure 3 reports theta recovery expressed as RMSE values. While the conclusions are in line with the bias results, patterns of design performances across routing probabilities are more pronounced. One notable pattern was observed with regard to the poorer performance of the NC methods when compared to the random and IRT-based methods for selecting the next optimal testlet for all three groups, even though the RMSEs were higher for the more extreme groups (i.e., low and high) as suggested by larger magnitudes of RMSEs for the four designs.

The bottom row of Figure 3 shows correlations between true and estimated person proficiency values. It was noted that high correlations were found mostly in the medium group and for the IRT-based and random methods, regardless of the routing probability. Most notably, for NC methods in low/high groups, correlations decreased as the routing probabilities moved away from 1.0 (optimal routing), with *L-S* design being the most affected (correlations ranged from mid-.50 at its highest to .40 at its lowest). The order (pattern of the performance) of the four studied designs also remained unchanged, with *S-L* yielding the highest correlations across routing probabilities in low and high groups, while *L-S* yielded the lowest correlations.

Section II. Item Exposure Rates and Parameter Recovery

In the current study, we were interested in examining designs that would yield the most even item exposure rates. For space reasons, we summarize exposure rates across the studied designs and conditions for different types of items. Specifically, in

Table 3 we report the average of exposures for items that are (albeit arbitrarily) divided into three groups: easy (items with generated b values less than –1), moderate (items with generated b values between –1 and +1), and difficult (items with generated *b* values greater than +1). This grouping is consistent with our person parameter criteria.

We note that Table 3 does not include items at the Core stage (e.g., the first 12 items that all simulees were administered, because their exposure rates were 1.00). Further, in order to gain insight into the exposures for various types of items (e.g., easier, or those more difficult according to their *b* value), we averaged across items with similar difficulty levels (presented in Table 3 as b level). Below, we highlight several interesting findings, first by looking across the different designs and then within each routing method. With respect to different study designs—albeit not surprisingly it was noted that when the design was balanced (*EQ* design) and when the routing method to select the next testlet was randomly selected—items of moderate difficulty were more exposed (.50) than either easier or more difficult items (at .36 and .27 exposure rates, respectively). Similarly, across the remaining four routing methods, items in the moderate range of difficulties were more exposed than items on either extreme (ranges from .49 to .61). However, an effect for suboptimal routing was observed such that as suboptimal probabilities increased (going from .00 to .20 to .30 for probabilities of 1.0, .80, and .70, respectively), item exposure rates for easier/difficulty items also tended to increase. This was the case for all conditions except for the IRT information method for difficult items, in which the exposure rates decreased, in contrast to what we would have expected.

Under the *S-L* design similar findings were observed for the random selection routing method as in the *EQ* design. That is, the moderate items were more exposed at .50 proportion than either easy or difficult items (.31 and .26, respectively). Additionally, as the probability of optimal routing decreased (from 1 to .7), exposure rates for moderate items decreased across other four routing methods as well. More so, the NC last testlet score method yielded the highest exposure rates for easy items under the *S-L* design, with ranges of .30 to .34. The highest item exposure rates for difficult items were observed for the IRT information method, ranging from .26 to .28; while the lowest item exposure rates for difficult items were observed for the NC cumulative score method, ranging from .12 to .21.

While similar patterns to those observed in the *EQ* and *S-L* designs were found in the *L-S* design, on average, easy and difficult items were exposed at higher rates than in the previous two designs. The highest exposure rates for easy items were found under NC last testlet score, which ranged from .39 to .41, and under IRT information method for difficult items with an exposure rate ranged of .31 to .37. Under the *S-L-S* design, exposure rates for easy or difficult items were all above .30 across all routing methods except NC cumulative score method and when applying NC last testlet score method at probability of 1 as exception.

Examining exposure rates within a particular routing method, we observed that when random selection was utilized, the *S-L* design yielded the lowest exposure rates for easy/difficult items at .31 and .26, respectively, while the *S-L-S* designs yielded the highest exposure rates in the range of .34 to .40. Comparing the classical approaches NC cumulative score and NC last testlet score, the latter yielded equally

Table 3
Average Item Exposure Rates (in Proportions) Across Routing Methods and Probabilities of Routing

	Routing to Next Testlet Methods											
	Random	Number-correct Cumulative Score			Number-correct Last Testlet Score			IRT EAP			IRT Information		
b level/prob	1	1	.8	.7	1	.8	.7	1	.8	.7	1	.8	.7
					Panel (a) *EQ* (12-12-12)								
Easy (22)	.36	.29	.32	.34	.32	.34	.36	.30	.32	.34	.26	.30	.32
Moderate (25)	.50	.61	.55	.53	.56	.52	.49	.60	.57	.54	.55	.54	.52
Difficult (13)	.27	.19	.23	.25	.23	.26	.29	.18	.21	.23	.34	.30	.29
					Panel (b) *S-L* (6-10-20)								
Easy (27)	.31	.23	.27	.29	.30	.32	.34	.22	.25	.28	.22	.25	.28
Moderate (32)	.50	.66	.59	.56	.56	.52	.48	.65	.59	.55	.57	.55	.53
Difficult (21)	.26	.12	.18	.21	.18	.23	.26	.16	.20	.23	.28	.26	.26
					Panel (c) *L-S* (20-10-6)								
Easy (15)	.40	.39	.40	.40	.39	.40	.41	.36	.37	.38	.36	.37	.38
Moderate (16)	.50	.54	.51	.51	.53	.51	.48	.57	.55	.53	.50	.51	.50
Difficult (7)	.29	.21	.25	.27	.24	.27	.30	.20	.23	.25	.37	.32	.31

(*Continued*)

Table 3
Continued

	Routing to Next Testlet Methods											
	Random	Number-correct Cumulative Score			Number-correct Last Testlet Score			IRT EAP			IRT Information		
b level/prob	1	1	.8	.7	1	.8	.7	1	.8	.7	1	.8	.7
						Panel (d) *S-L-S* (6-20-10)							
Easy (24)	.40	.37	.38	.39	.41	.41	.42	.31	.35	.36	.30	.33	.36
Moderate (30)	.50	.59	.55	.54	.54	.51	.49	.59	.56	.53	.57	.54	.53
Difficult (16)	.34	.21	.27	.29	.25	.30	.33	.30	.32	.33	.37	.35	.35

Notes. b level Easy = values in cells represent average proportions of items with *b* value less than –1; Moderate indicates average proportions of items with *b* values between –1 and +1; Difficult indicates average proportions of items with *b* values > +1. Numbers in () represent the number of items that belong to a particular b level, of which the averages were taken and reported. Prob refers to the probability of routing correctly to the next testlet such that 1 means that the next testlet is always correctly selected, while probabilities of .8 and .7 represent conditions in which suboptimal routing is modeled in 20% and 30% of examinees. Each panel represents a different design used in the study: *EQ* - Design 1 (equal) had 12 items in each of the three testlets; *S-L* - Design 2 (short-to-long) had six items in the Core testlet, followed by 10 and 20 items in subsequent testlets; *L-S* - Design 3 (long-to-short) had 20 items in the Core testlet, followed by 10 and 6 items in subsequent testlets; *S-L-S* - Design 4 (short-long-short) had six items in the Core testlet, followed by 20 and 10 items in subsequent testlets.

high or higher exposure rates for easy and difficult items, and lower exposure rates for moderate items, regardless of the design. The most extreme differences among the rates between the two classical approaches were found in exposure rates under the *S-L* design for moderate items. In particular, under a routing probability of 1, exposure rates for NC cumulative score method were .66 (which was also the highest exposure rates reported for any item type across any method and design) compared to .56 for its NC last testlet score counterpart.

Examining performance of the IRT-based routing methods, EAP ability and information, a pattern was noted. For difficult items, using IRT information, items were exposed more frequently on average than using EAP ability estimate as a routing method. However, the opposite was true for the easy items, which were exposed generally at equal or higher rates using EAP ability estimates compared to IRT information. The exposure rates for easy items were equivalent across the two methods and across all three routing probabilities between EAP ability and IRT information under the *S-L* design (.22, .25, and .28) and the *L-S* design (.36, .37, and .38), respectively.

A brief remark regarding the exposure rates concerns the idea of suboptimal routing. As noted above, the probability of routing had a major impact on item exposures for all items. This was not surprising as the study design was meant to impact item exposure such that some people were purposefully misrouted with the goal of evening out exposure rates. However, different routing methods were impacted in different ways. When it came to moderate items, in all but one case, an increase in misrouting (i.e., probability of selection of optimal testlet decreases from 1 to .70) yielded lower exposure rates of items that ranged from –1 to +1 (moderate items). One exception was noted under IRT information in the *L-S* design where the exposure rates for moderate items roughly maintained at .50 (specifically, exposure rates at 1.00, .80, and .70 probability were .50, .51, and .50, respectively). This irregular pattern of decreased exposure rate across routing probabilities less than 1 was also noted across all designs (*EQ*, *S-L*, *L-S*, and *S-L-S*) for difficult items under IRT information, but it was not noted for easy items where exposure rates either remained the same or increased across all methods and designs.

Recovery of item difficulty (Figure 4) and item discrimination (Figure 5) parameters varied depending on the type of items. Namely, in general, items that were recovered with the highest level of precision (i.e., smallest biases/RMSEs, highest correlations) were those for items in the middle of distribution—that is, items considered moderate in difficulty regardless of the probability routing levels or designs (with a couple of exceptions in *L-S* designs). More so, as Figure 4 suggests, bias for easy items ranged from –.01 to .06, with recovery being influenced by the design choice (note the irregular scatter of the markers within any one routing method, in particular for the NC routing methods). Similar, and perhaps a slightly more pronounced reverse patterns were found for difficult items, where unlike for the easier items, bias tended to be near zero or negative.

RMSE values for item difficulty recovery provide further insight into the irregularity of the difficulty recovery across the designs and studied conditions. Namely, in some instances, *S-L-S* design yielded lower RMSEs (IRT-based routing methods and NC last testlet for easy items), while in others it yielded one of the highest RMSE values (difficult items under NC cumulative testlet score with routing probability of 1).

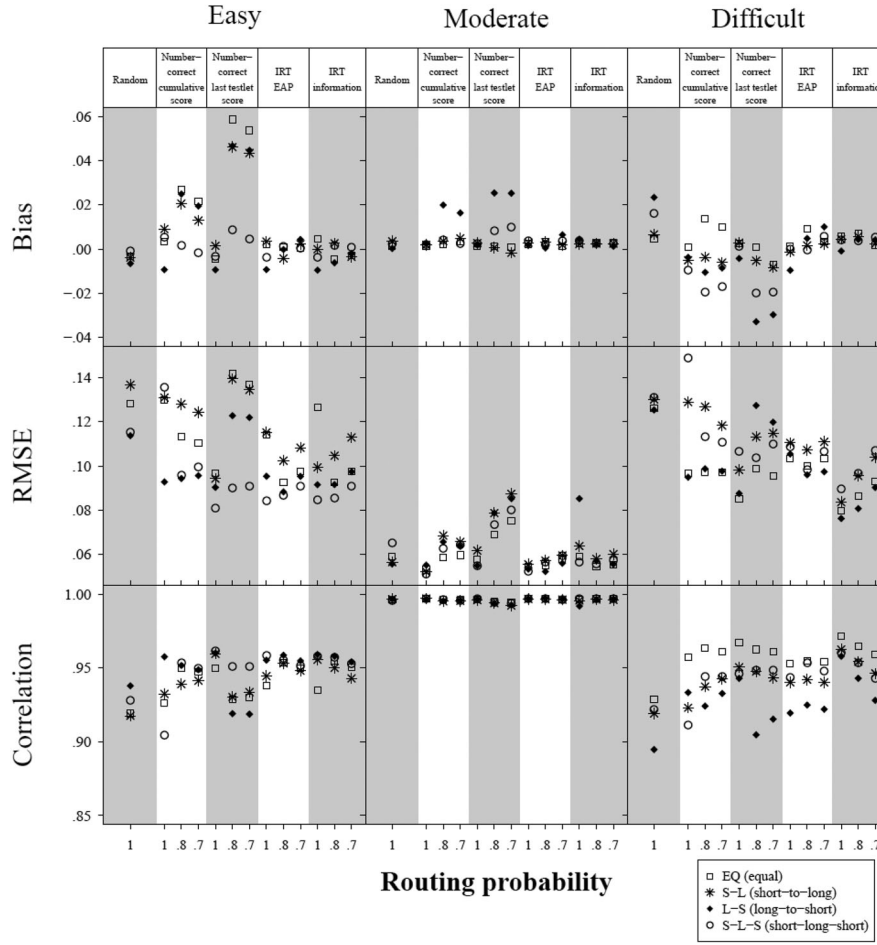

Figure 4. Item difficulty parameter recovery rates reported as bias, RMSE, and correlations for easy, moderate, and difficulty items.

The inconsistency of the methods across the four designs was most clearly noted using the RMSE values as an outcome to evaluate the recovery of item difficulty parameters under studied conditions. It was thus not surprising that when examining correlations, correlations of nearly 1.0 were found for moderate items, while lower correlations were found for easy and difficult items. However, we note that almost all correlations were .90 or higher (with one exception for *L-S* design under random method at routing probability of 1 for difficult items, where the correlation dipped just below .90).

The recovery of item discriminations varied across the routing methods for all three types of items. While the random routing method yielded, on average, the least amount of bias across the four studied designs (in particular for easy items), discriminations of moderate items were recovered most similarly across the four studied design and IRT-based routing methods yielded on average the lowest bias.

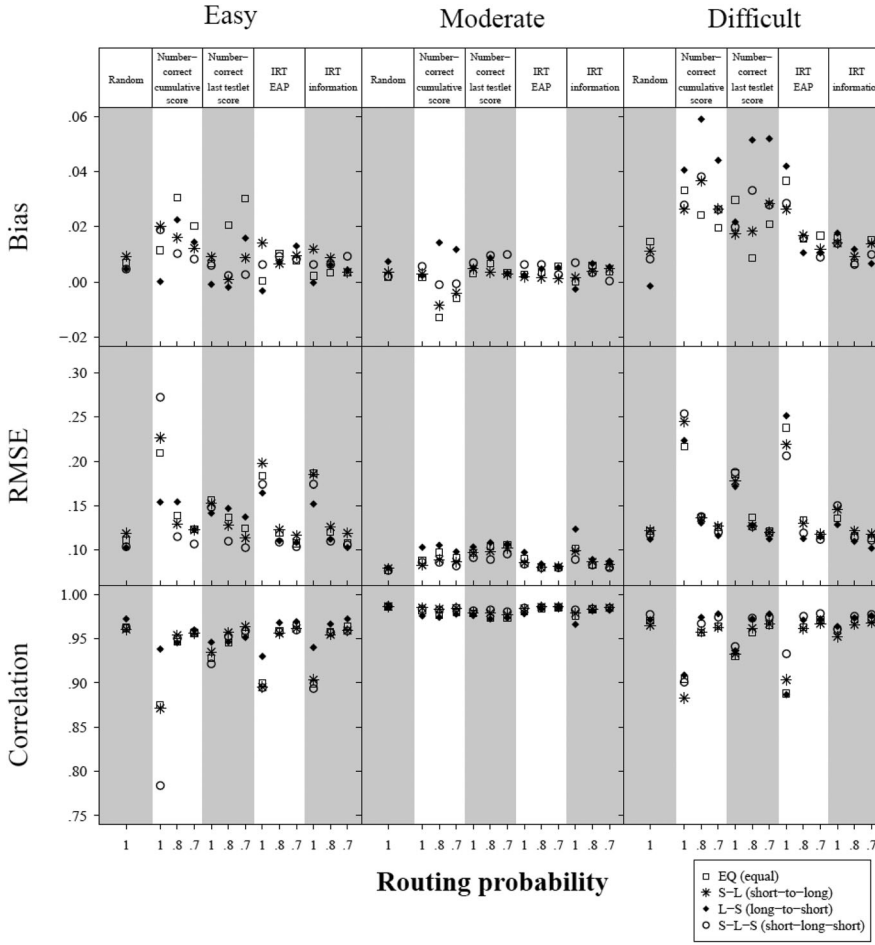

Figure 5. Item discrimination parameter recovery rates reported as bias, RMSE, and correlations for easy, moderate, and difficulty items.

It was also noted that item discriminations were recovered to a varying degree based on the routing probabilities and routing methods, with no one routing method outperforming the others or any one design yielding the lowest bias across different items types. For example, the *E-Q* design was impacted more (i.e., largest differences in bias) by routing probabilities under the NC methods for easy items (yielding as large or larger bias than other designs). However, for difficult items, the results were opposite—under the *E-Q* design, routing probabilities away from 1.0 yielded less bias for those same routing methods. Similarly, for difficult items and the two NC routing methods, *L-S* design produced the highest bias for routing probabilities of .8 and .7, but yielded the lowest bias for routing probabilities of 1.0 for the easy items.

RMSEs (as shown in the second row of Figure 5) further supported conclusions that across the different item types, no one routing method consistently outperformed

the others, nor did one design yield the lowest RMSEs across the studied conditions. It was noted that the moderate items' discriminations were most successfully recovered under all five routing methods. Different ordering of the designs (i.e., when a design yielded lowest RMSEs) was more pronounced for the easy and difficult items, in particular when routing probabilities were set at 1.0. Additionally, for difficult items and routing probabilities of .8 or .7, item discriminations were recovered very similarly among the four designs for all five routing methods. This was less obvious for the easy items' recovery of discrimination, which RMSEs tended to be slightly higher than in other item groups.

Correlations were generally high, with most of them above .90 across the easy/moderate/difficult items, with only a few instances where the correlations between the generated and estimated item discriminations were below .90 (last row of Figure 5). Most notably, the lowest correlations were obtained in conditions under routing probability of 1.0 for *S-L-S*, *S-L*, and *E-Q* designs for easy and difficult items. Out of the five routing methods, random generally yielded as high or higher correlations across all designs.

Discussion

ILSAs' move from paper-and-pencil administration to a computerized platform affords testing organizations a number of advantages, including the possibility of implementing an adaptive test. This move is particularly important in low-performing countries where average student proficiency is well below the international average. Without some type of adaptive testing the vast majority of students in these systems receive questions that are too difficult, offering little opportunity to engage with items and potentially resulting in biased achievement estimates for the lowest performers (Rutkowski, Rutkowski, & Liaw, 2018). In response, PISA is implementing a multistage test design for the 2018 cycle. With the promise of MST, a number of open questions remain, particularly in the unique context of ILSA.

To begin answering some of these research questions, we used a simulation study to examine several design possibilities, including testlet length and routing procedures. We considered these design features in terms of item and person parameter recovery as well as item exposure rates. Our findings suggest that no single approach is best for all purposes. We summarize this point subsequently. In terms of mean person parameter recovery, IRT scoring (via either Fisher information or preliminary EAP estimates) outperformed classical NC methods. This was true in terms of bias, RMSE, and correlations between generating and estimated proficiency. In spite of this well-known performance advantage, we note that from an operational perspective, NC methods offer a much simpler algorithm that is easier to implement. To that end, NC methods exhibited only slightly worse bias, especially if suboptimal probability routing is used with cumulative scoring for routing to the third stage. Alternatively, if only the last testlet is considered for routing, a design with fewer questions in the first (routing) stage is better at recovering mean proficiency, with lower RMSE values and higher generating-estimated proficiency correlations, compared to all other testlet length designs.

We also considered the way in which design and scoring choices impacted item exposure rates and recovery. With respect to exposure rates, as expected, suboptimal routing produced the most even item exposure across scoring methods and testlet lengths. Further, in nearly every condition with a routing probability less than 1, moderately difficult items were exposed to about 50% to 60% of examinees, while easy and difficult items were exposed to about 30% to 40% of examinees. We found, however, that IRT routing and a short routing test produced lower exposure rates, especially for difficult items. Our primary interest in exposure rates was driven by an interest in producing stable item parameters, which we turn to next.

Item difficulties were generally recovered well, with bias ranging from –.04 to .06 across the varying routing methods and designs. Items considered moderate in difficulty were recovered best, and most consistently across the four designs. Easy and difficult items tended to be biased in terms of recovery, and in most occasions the NC-based routing methods tended to be less precise in difficulty parameter recovery. Similarly, items' discriminations were generally recovered well (with magnitude of bias of .04 or lower for most conditions), especially for items considered of moderate difficulty. In general, random and IRT-based routing methods tended to outperform the NC methods, although this pattern did not hold across the four studied designs. In other words, the length of the testlets yielded varying degrees of precision in item discrimination recovery within a routing method.

In summary, this article provides some practical guidance for testing organizations, as they begin and/or continue to implement adaptive designs. As a simulation study, this project has limitations. First, although we considered several design and administration conditions and we modeled our simulation after empirically observed conditions, other options exist. For example, we considered a single population with no model misspecification. And further research with many groups and, especially, violations of the measurement invariance assumption, are needed for a fuller picture. Similarly, an issue of item drift should be considered, with focus on establishing ways by the programs to prevent item parameter from drifts. Nevertheless, our findings point to the fact that no single design will meet all goals. Rather, a careful evaluation of trade-offs should be made for any ILSA design.

Acknowledgments

This research was supported in part by a grant from the Norwegian Research Council under the FINNUT program (Grant Number 255246).

Notes

[1]In the MST literature, panels represent a test form in testing. However, there could exist multiple different forms within the same panel structure. In the current study, we utilize a single panel as represented in Figure 2, but we manipulate different test lengths, thus creating multiple forms.

[2]In their edited volume, Yan, von Davier, and Lewis (2014) provide an in depth research regarding important issues related to computerized multistage testing, in which authors and co-contributors discuss an array of MST designs and applications.

[3]Weissman (2014) discusses in depth two routing rules that these two broad approaches encompass: (a) a static routing rule(s) such as NC, and (b) dynamic routing rules. Within each type of routing rules, several decisions ought to be made with respect to administration. For example, when using a static routing rule, one ought to determine a threshold score that would apply to a group of examinees, whereas under dynamic routing rule different algorithms can be used in real time to make routing decisions (i.e., focus is on an individual test taker).

[4]The *mst* function was custom modified by Magis et al. (2018) to allow for probabilistic routing element.

[5]These choices of fixed factors are not unusual, albeit there is a wide range of sample sizes across studies. For example, Hambleton and Xing (2006) sampled 5,000 scores from a standard normal distribution for their 1-3-3 MST design, with 20 items per testlet. Chuah, Drasgow, and Luecht (2006) suggested that as few as 300 examinees per item might be sufficient in MST design for accurate item parameter estimation, whereas S. Kim et al. (2015) simulated 2,000 simulees at each of the 41 quadrature points across the continuum for a total sample size of 82,000.

[6]Star represents the *S-L* design that had six items in the Core testlet, followed by 10 and 20 items in subsequent testlets; *L-S* design (diamond) had 20 items in the Core testlet, followed by 10 and 6 items in subsequent testlets; and *S-L-S* (circle) had six items in the Core testlet, followed by 20 and 10 items in subsequent testlets.

References

Armstrong, A. (2002). *Routing rules for multiple-form structures* (LSAC Research Report Series No. 02–08). Newtown, PA: Law School Admission Council. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.310.3575&rep=rep1&type=pdf

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chen, H., Yamamoto, K., & von Davier, M. (2014). Controlling multistage testing exposure rates in international large-scale assessments. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 391–409). Boca Raton, FL: CRC Press.

Chuah, S. C., Drasgow, F., & Luecht, R. M. (2006). How big is big enough? Sample size requirements for CAST item parameter estimation. *Applied Measurement in Education*, *19*, 241–255.

Educational Testing Service. (2016). *PISA 2018 integrated design*. Princeton, NJ: Author. Retrieved from https://www.oecd.org/pisa/pisaproducts/PISA-2018-INTEGRATED-DESIGN.pdf

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

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice*, *26*(2), 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*, 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 meeting of the National Council on Measurement in Education, Atlanta, GA.

Kim, S., Moses, T., & Yoo, H. H. (2015). A comparison of IRT proficiency estimation methods under adaptive multistage testing: A comparison of IRT proficiency estimation methods. *Journal of Educational Measurement*, *52*, 70–79.

Kirsch, I., & Lennon, M. L. (2017). PIAAC: A new design for a new era. *Large-Scale Assessments in Education*, *5*(1). https://doi.org/10.1186/s40536-017-0046-6

Kolen, M. J., & Tong, Y. (2010). Psychometric properties of IRT proficiency estimates. *Educational Measurement: Issues and Practice*, *29*(3), 8–14.

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

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

Magis, D., Yan, D., & von Davier, A. (2018). mstR: Procedures to generate patterns under multistage testing (R package version 1.2). Retrieved from https://CRAN.R-project.org/package=mstR

Melican, G. J., Breithaupt, K., & Zhang, Y. (2009). Designing and implementing a multistage adaptive test: The uniform CPA exam. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 167–189). New York, NY: Springer.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational and Behavioral Statistics*, *17*(2), 131–154.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 international results in mathematics*. Boston, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College. Retrieved from http://timssandpirls.bc.edu/timss2015/international-results/timss-2015/mathematics/student-achievement/

OECD. (2010). *PISA Computer-based assessment of student skills in science*. Paris, France: OECD Publishing. Retrieved from http://www.oecd.org/education/school/programmeforinternationalstudentassessmentpisa/pisacomputer-basedassessmentof-studentskillsinscience.htm

OECD. (2014). *PISA 2012 technical report*. Paris, France: OECD Publishing.

OECD. (2017). *PISA 2015 technical report*. Paris, France: OECD Publishing. Retrieved from http://www.oecd.org/pisa/data/2015-technical-report/

R Development Core Team. (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/

Robin, F., Manfred, S., & Liang, L. (2014). The multistage test implementation of the GRE revised general test. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 325–341). Boca Raton, FL: CRC Press.

Rutkowski, D., Rutkowski, L., & Liaw, Y.-L. (2018). Measuring widening proficiency differences in international assessments: Are current approaches enough? *Educational Measurement: Issues and Practice*, *37*(4), 40–48.

Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, *23*(1), 57–75.

Weissman, A. (2014). IRT-based multistage testing. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 153–168). Boca Raton, FL: CRC Press.

Weissman, A., Belov, D., & Armstrong, A. (2007). *Information-based versus number-correct routing in multistage classification tests* (LSAC Research Report Series No. 07–05). Newtown, PA: Law School Admission Council. Retrieved from https://www.lsac.org/docs/default-source/research-(lsac-resources)/rr-07-05.pdf

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

Yan, D., Lewis, C., & von Davier, A. A. (2014a). Multistage test design and scoring with small samples. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 303–324). Boca Raton, FL: CRC Press.

Yan, D., Lewis, C., & von Davier, A. A. (2014b). Overview of computerized multistage tests. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 3–20). Boca Raton, FL: CRC Press.

Zheng, Y., Wang, C., Culbertson, M. J., & Chang, H.-H. (2014). Overview of test assembly methods in multistage testing. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 87–99). Boca Raton, FL: CRC Press.

Authors

DUBRAVKA SVETINA is Associate Professor of Inquiry Methodology, School of Education, Indiana University, 201 N. Rose Avenue, Bloomington, IN 47405; dsvetina@indiana.edu. Her primary research interests include educational and psychological measurement, (international) large-scale assessment, item response theory, measurement invariance, and psychometric modeling (e.g., Bayesian and cognitive diagnostic models).

YUAN-LING LIAW is Postdoctoral Researcher at the Centre for Educational Measurement (CEMO), Faculty of Educational Sciences, University of Oslo, P.O Box 1161, Blindern, 0318 Oslo, Norway; y.l.liaw@cemo.uio.no. Yuan-Ling's primary research focuses on practical applications of item response theory, with greatest emphasis on international large-scale assessment. These include test fairness, differential item functioning, and ability estimation.

LESLIE RUTKOWSKI is Associate Professor of Inquiry Methodology, School of Education, Indiana University, 201 N. Rose Ave, Bloomington, IN 47405; lrutkows@iu.edu. Her primary research interests are in latent variable modeling, especially models that pertain to cross-cultural measurement and international comparisons among heterogeneous populations.

DAVID RUTKOWSKI is Associate Professor of Educational Policy, School of Education, Indiana University, 201 N. Rose Ave, Bloomington, IN 47405; drutkows@iu.edu. His primary research interests are in educational policy and large-scale assessment, focusing on cross-cultural measurement and international comparisons among heterogeneous populations.