Valdivia Medinaceli *et al.*
Large-scale Assessments in Education (2023) 11:22
https://doi.org/10.1186/s40536-023-00165-9

Large-scale Assessments in Education

METHODOLOGY Open Access

Effects of DIF in MST routing in ILSAs

Montserrat Valdivia Medinaceli[1*], Leslie Rutkowski[1], Dubravka Svetina Valdivia[1] and David Rutkowski[1]

*Correspondence:
mbvaldiv@iu.edu

[1] Counseling and Educational Psychology, Indiana University Bloomington, 201 N. Rose Ave., Bloomington, IN 47405, USA

Abstract

The advance and access to technology have allowed for the implementation of adaptive testing in international large-scale assessments (ILSAs). Multistage testing (MST) in ILSAs offers opportunities and advantages compared to linear testing. However, when dozens of highly heterogeneous systems participate in ILSAs, the sources of heterogeneity pose challenges to any cross-cultural measurement endeavor. With the recent implementation of adaptive designs in ILSAs, little is known about differential item functioning (DIF) effects, mainly when an MST design is used in an ILSA context. Through a simulation study grounded on the empirical basis of the Trends in International Mathematics and Science Study (TIMSS) data, this paper examines the impact of DIF in MST routing in ILSA under different routing strategies. Results showed that Merit routing is highly accurate even when the amount and magnitude of DIF is high, whereas suboptimal routing showed poor accuracy across DIF conditions. As expected, Merit routing has better proficiency recovery parameters than a suboptimal routing mechanism. Implications and recommendations for test developers are included in the discussion section.

Keywords: Differential item functioning, Multistage testing, International large-scale assessment

Introduction

In international large-scale assessments (ILSAs), adaptive testing offers advantages over traditional (e.g., linear) test administration methods. Advantages of adaptive testing include efficiency gains, improved precision about proficiency estimates, especially at extreme ends of the proficiency continuum (Betz & Weiss, 1974; Hambleton & Swaminathan, 1985; Lord, 1980; Wainer et al., 1992), and a better test-taking experience (Yan et al., 2016). Dozens of highly heterogeneous systems from varied cultural, geographic, linguistic, and socioeconomic backgrounds participate in ILSAs. For example, if we select two participants from the Trends in International Mathematics and Science Study (TIMSS) 2019 and compare their gross domestic product (GDP) per capita, such as Pakistan ($1202) and Ireland ($100,172; World Bank, 2021), we observe a stark difference in economic resources which undoubtedly affect the allocation of resources in each respective educational system. Besides differences in participants' economic resources, geographic, cultural, and linguistic differences pose challenges to the quality of any cross-cultural measurement endeavor. As the participation of heterogeneous populations in ILSAs grows, this challenge can also be observed in a broadening of the proficiency continuum (Rutkowski et al., 2018). Therefore,

the possibility of adaptive testing—wherein the difficulty of items or groups of items are tailored to the proficiency of the examinee—is appealing for ILSAs and particularly helpful for measuring proficiency at the ends of the proficiency continuum.

One of the most important methodological challenges in cross-cultural measurement is the assumption that item characteristics are invariant across groups. In other words, the models used to estimate proficiency or other latent attributes assume that these constructs are understood and measured equivalently across countries (Lord, 1980; Millsap, 2011). However, violations of this assumption—commonly known as differential item functioning (DIF)—occur to varying degrees in the ILSA context (Ercikan, 1998; Grisay & Monseur, 2007; Rutkowski & Rutkowski, 2013). Although the impact of DIF is reasonably well understood for linear ILSAs (Oliveri & von Davier, 2011; Rutkowski et al., 2016; Svetina & Rutkowski, 2014), much less is known about the impact of DIF in an adaptive ILSA context. We take up this issue in the current paper. Importantly, MST has been implemented in the Programme for the International Assessment of Adult Competencies 2011 (PIAAC; Organization for Economic Cooperation and Development, 2013; Kirsch & Lennon, 2017; Kirsch et al., 2020) and in the Programme for International Student Assessment's (PISA) 2018 reading domain (Educational Testing Service, 2016; Yamamoto et al., 2018; Yamamoto et al., 2019). Moreover, group adaptive testing design was implemented in the Progress in International Reading Literacy Study (PIRLS) 2021 (Mullis & Martin, 2019) and in TIMSS 2023 (Yin & Foy, 2021). Group adaptive testing design is based on Pohl's (2013) longitudinal MST design, where group (i.e., countries) scores from a previous administration (i.e., PIRLS 2016) are used to provide more or less difficult sets of items to that group in the current test administration (i.e., PIRLS 2021). As ILSAs transition into adaptive testing, questions regarding the potential implications of DIF in routing decisions are extremely relevant in the considerations for the unbiased item and proficiency estimation.

In general, test adaptability occurs either at the item level or the item cluster level. The former is referred to as a computerized adaptive test (CAT), while the latter is referred to as a multistage test (MST). MST is often a more desirable operational choice (Melican et al., 2009) because, compared to linear fixed-length tests, MST has shown greater testing efficiency and increased accuracy in proficiency estimates (Jodoin et al., 2006; Kim & Plake, 1993). Furthermore, Luo and Kim (2018) noted that MST provides several practical advantages over CAT. For example, one advantage over CAT is a priori knowledge of psychometric and content properties of all possible test forms. The main advantage is that MST designs can be constructed prior to administration, providing a more efficient approach to dealing with complex test constraints and minimizing computing complexity while providing flexibility for the test taker to review and revise responses in the same stage of the assessment. In what follows, we offer a brief overview of an MST design followed by a discussion of how DIF might prove especially challenging in an ILSA MST setting.

Background

MST is an algorithm-based approach that assigns a preassembled set of items to test-takers following a sequential design. In other words, a test-taker gets sets of items (i.e., modules) with a difficulty level that better fits their proficiency level (Yan et al., 2016). MST is linear in that the adaptation happens at the module—sets of items—rather than at the item level as in a fully CAT design. An MST design usually features panels, of

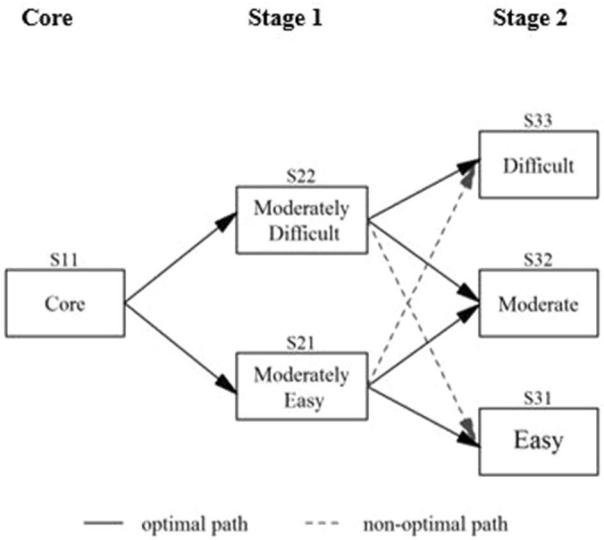

Figure 1 An example of a 1–2–3 three-stage MST design used in the study

which there can be one or many. Each panel is comprised of modules that are arranged into sequential stages, from 1 to *n*. The first module (Core) is the starting point for all examinees. Figure 1 illustrates a single panel and shows the various paths that an examinee may take. According to this design, each examinee receives three blocks of non-overlapping items. An MST design adaptation depends on an examinee's performance and the routing rule. Routing rules are simpler to use for estimating performance and are determined prior to the MST administration (Weissman, 2016); hence, decisions concerning routing to the next module are a priori identified. In Fig. 1, MST adaptive paths can be optimal or non-optimal. Optimal paths are those identified a priori (i.e., Core-Difficult-Difficult, Core-Difficult-Moderate, Core-Easy-Moderate, Core-Easy-Easy), and non-optimal are a result of the misrouting mechanism (i.e., Core-Difficult-Easy, and Core-Easy-Difficult). For example, in Fig. 1, a student routed out of the Core module into a moderately difficult Stage 1 module should be routed to either a moderate or difficult Stage 2 module based on their performance on Stage 1, the Core, or both. The described design is just one of many possibilities in the MST framework.

MST offers opportunities and poses challenges in the ILSA context. With the participation of more developing economies in ILSAs, the spectrum of proficiency distribution has been extended. Assessments that have been created for developed countries—such as in PISA, where the original participants were overwhelmingly country members of the OECD—, now have more heterogeneous participants, making the design of a test that can measure proficiency well across the spectrum more challenging (Rutkowski & Rutkowski, 2021). The adaptive component of an MST assessment design allows for better measurement of the proficiency of participating countries across the continuum in ILSAs. For example, gains in proficiency estimation precision exist for an MST design in comparison to a linear test design (Jodoin et al., 2006), specifically in PIAAC, where MST was able to obtain the same test information with 13–32% fewer items than a linear test (OECD, 2013). Moreover, MST allows control over certain ILSA design features in

its implementation, such as content balance within modules (Yamamoto et al., 2018), and limits the exposure of items across the design (Svetina et al., 2019). For all its strengths, MST does have some challenges in its implementation. For example, not all heterogeneous populations and subpopulations have access to or familiarity with using computers, which could present itself as a DIF problem when MST designs are implemented in ILSA.

ILSA data are primarily used to compare countries, making it essential to assure measurement invariance, which can be examined at a scale level (i.e., multigroup CFA) or item level (i.e., DIF). Measurement invariance, or equivalence, is a psychometric quality that states that the results of a test are equivalent across different groups (Meredith, 1993; Millsap, 2011). Obtaining measurement invariance allows for fair and valid comparisons of test results across groups. Differential item functioning (DIF) is defined as an expected item score difference across multiple groups after the scales and scores are calibrated and linked on the same metric (Holland & Wainer, 1993; Lord, 1980). DIF is a threat to ILSA's uses, interpretation, and inferences validity because the presence of DIF threatens score equivalence (Ackerman, 1992; Millsap & Everson, 1993), reducing the fairness of cross-cultural comparisons and inferences (Zwick et al., 1997).

Little is known about the effects of DIF when an MST design is used in an ILSA context. Previous research on the effects of DIF in ILSAs suggests that, in a linear design with random assignment of booklets/forms, DIF affects countries' rankings based on their proficiency estimates (Oliveri & von Davier, 2011). Under an MST design, test-takers would instead receive a module with a difficulty level matched to their proficiency level after taking the core module. The items within each module affect the final proficiency more than when randomly assigned. Therefore, DIF on any item may have more consequential results at the test-taker level (Gierl et al., 2013; Zwick, 2010; Zwick & Bridgeman, 2016). DIF could negatively affect the adaptive procedures and have detrimental consequences on the examinees' proficiency estimates (Zwick, 2010). Adaptive tests are more sensitive to the effects of DIF than linear tests (Steinberg et al., 2000). More specifically, CAT is more resilient in their proficiency estimation than MST when DIF is present (Kara & Dogan, 2022). However, all these studies aimed at the individual test taker level rather than the group level, making further investigation in this context relevant.

The MST design could be even more complicated if a classification matrix is used, as in PISA 2018, where students were assigned to subsequent modules based on different routing probabilities. As examinees are routed based on their score in previous modules, the presence of DIF can potentially drive routing decisions, which could ultimately influence proficiency estimates at a population level. That is, if an item or group of items are, for instance, differentially more difficult for a group of students, a risk of routing to an easier module exists, which could lead to lower proficiency estimates than expected. The opposite is also possible—differentially easier items could route students to more difficult modules, risking bias in proficiency estimates.

The degree to which routing errors and associated cascading impacts occur because of DIF in an MST design is an understudied area. Further, the unique context of ILSAs, where inferences are focused at the population level (von Davier & Sinharay, 2014), and complex methods are used to estimate proficiency (Mislevy et al., 1992; von Davier &

Sinharay, 2014), raises important questions about DIF and MST. ILSAs complex population estimation consists of a latent regression modeling approach to overcome challenges associated with examinees getting booklets assigned at random (Mislevy et al., 1992). The latent regression modeling approach treats the examinee's proficiency as missing data. A conditional model analogous to a multiple imputation approach is used to obtain the missing data. In other words, an imputation model is constructed (Rubin, 1976), imputing proficiency for all examinees and drawing plausible values (Mislevy, 1991) for each individual respondent (von Davier & Sinharay, 2014). In this paper, we examined the impact of DIF on routing decisions, item parameter estimates, and the consequences that biased item parameters might have on population-level proficiency distribution estimates. Also, we examined the degree to which DIF *location* matters. That is, we investigated whether differential impacts depend on where DIF occurs in the early versus later stages of the assessment.

Methods

To address the research questions, we conducted a Monte Carlo simulation study using the software *R* (R Core Team, 2022). To ground our study in an empirical setting, we incorporated the TIMSS 2015 item and proficiency parameters for an eighth-grade mathematics assessment in nine representative countries. We note that although TIMSS does not currently use a traditional MST design, our findings should be reasonably generalizable to ILSA settings.

Fixed factors

This study considered a fixed sample size of 4000 individuals per country, with normally distributed generating proficiency distributions $\sim N(\mu, \sigma)$ based on nine TIMSS 2015 participating countries. Although overall sample sizes vary for participating countries, participating countries sampled between 3000 and 13,000 in TIMSS 2015. Most participants sampled 4000 examinees per country; therefore, a sample of 4000 simulees aligned with operational procedures. Additionally, an equal sample allowed for equal weight to each participant country in the calibration of items. The selected nine countries corresponded to the countries with the highest, medium, and lowest proficiencies according to TIMSS 2015 eighth-grade mathematics results. The generated mean proficiency parameters were obtained by standardizing mean ($\mu = \frac{Mean-500}{100}$) scores and standard deviations ($\sigma = SD/100$) reported by TIMSS 2015 (see Table 1).

Our simulation features a single MST panel with a 1–2–3 design, as seen in Fig. 1. Classical Test Theory (CTT) routing is used after a student has completed each module. Like PISA 2018's MST routing decision (Yamamoto et al., 2018), the routing module selection method is based on the total number of correct, automatically scored items on a given module. Moreover, we used a routing method that allows for the selection of subsequent modules using different routing probabilities. In ILSAs, probability routing prevents module over-exposure and allows for controlling content balance (Yamamoto et al., 2019). Including routing probabilities ensures that in a situation where low-performing countries will more frequently see easy items and high-performing countries will more frequently see hard items, those item parameter estimates are based on a representative sample of all test takers. The total number of items for each student in a panel

Table 1 Descriptive statistics of proficiency parameters (N = 4000 for each country)

	Country	*M*	*SD*	*μ*	*σ*
High performers	Singapore	621	82	1.21	0.82
	Republic of Korea	606	85	1.06	0.85
	Chinese Taipei	599	97	0.99	0.97
Medium performers	Sweden	501	72	0.01	0.72
	Malta	494	88	− 0.06	0.88
	Malaysia	465	87	− 0.35	0.87
Low performers	Morocco	384	80	− 1.16	0.8
	South Africa	372	87	− 1.28	0.87
	Saudi Arabia	368	86	− 1.32	0.86

Note: The Mean and Standard Deviation for each participant country in TIMSS 2015 was retrieved from http://timssandpirls.bc.edu/timss2015/international-results/timss-2015/mathematics/student-achievement/distribution-of-mathematics-achievement/

SD standard deviation, *N* sample size, μ mean proficiency parameter, σ proficiency parameter standard deviation

was 36, with an equal number of items per module within each stage. This resulted in 12 items per module and was consistent with previous studies (Verschoor & Eggen, 2014). Furthermore, we limited the generated data to dichotomous responses and uniform DIF to make the study manageable and avoid confounding factors. Previous studies reported that uniform DIF is more prevalent in operational settings (i.e., Joo et al., in-print). To study the effect of DIF on routing accuracy, we induced uniform DIF (Hanson, 1998), leading to a different item difficulty.

Manipulated factors

We manipulated several factors in this study, including the magnitude of DIF (small, medium, and large), the number of items with DIF within a module (one, three, and five), DIF location in the MST design (Core, Stage 1, or Stage 2), DIF country (low, medium, and high), and routing probabilities (Merit, M + PM, and Random). We provide details and rationales for each manipulated factor below.

DIF magnitude

Three levels of DIF magnitudes were manipulated in this study: small, medium, large. We were interested in observing if the magnitude of DIF plays a role in influencing routing, item parameters, and proficiency estimation and whether the location of DIF items (early or late in the MST design) mattered. A two-step process was conducted through empirical analysis of the TIMSS 2015 mathematical items to examine the DIF magnitude. Following operational practices (Martin et al., 2016), a 2-PL multigroup IRT model was fitted to the empirical pooled sample of the data (all participant countries) to estimate international item parameters. Then we estimated the proficiency for each country via a multigroup latent regression model, fixing the item parameters to the item estimates obtained from the previous step. The scale was set by fixing the first group to a standard normal distribution with $M = 0$ and $SD = 1$. Lastly, item-by-group parameters were estimated using a 2-PL IRT model. The scale was set to estimate group-specific item parameters by fixing each country's estimated mean proficiency. The estimated difference between group-specific and pooled sample item parameters resulted in the DIF

magnitude for difficulty parameters. Of 215 cognitive items, 12 had DIF more extreme than the latent trait continuum (± 3) for at least one country; these items were excluded. The results showed a large range of DIF magnitudes. We obtained the average absolute value of the bias in the difficulty parameter across all countries. Next, we selected the 25th, 75th, and 95th percentiles of estimated DIF magnitudes. We observed DIF values for the difficulty parameter of 0.30 (small DIF), 0.60 (moderate DIF), and 1.40 (large DIF). According to the condition, these magnitudes were added to the difficulty parameters for the respective focal group (see DIF country).

DIF amount

For this study, a baseline and three DIF amounts (i.e., number of items per module) were simulated based on the RMSD analysis of TIMSS 2015 results previously described. The proportion of DIF flagged items over the total number of items was evaluated. The first quartile (10%), third quartile (22%), and maximum (43%) were chosen to simulate the number of DIF items based on the TIMSS 2015 pool of items. In this simulation study, out of the 12 items, the DIF amounts were simulated by randomly inducing DIF to 1, 3, and 5 items per module, respectively.

DIF location

To understand how DIF location in the MST design impacts routing, we isolated DIF to only appear in one stage at a time. Stages with more than one module (i.e., Stages 1 and 2) would have the same DIF magnitude and amount in all modules in that stage. In that way, we can study the effects of DIF on routing accuracy and proficiency parameter recovery.

DIF country

We investigated three levels of group proficiency distributions for the DIF group: low, medium, and high-performing countries. We were particularly interested in evaluating the effects of DIF in MST routing when different countries experience bias and comparing if there are any differences when DIF is present for countries at the end or middle of the proficiency continuum. We induced DIF in low and medium performing countries adding to the difficulty parameters, while in high performing countries we induced DIF by subtracting from the difficulty parameters of the items selected. This decision was based on results from the empirical analysis, where DIF was mostly negative for low and medium performing countries, and positive for high performing countries.

Routing probabilities

Previous research showed that routing probabilities are necessary to ensure even item exposure across highly diverse participating countries (Svetina et al., 2019). To examine the effects of DIF on routing scenarios, we considered three routing probabilities: Merit only, Merit with probabilistic misrouting (M + PM), and Random routing. Merit only routing is fully determined by the examinee's performance, implying that all test takers will be routed to the module that best matches their proficiency, with a probability equal to 1. In M + PM routing, examinees have a 30% probability of *misrouting*, regardless of proficiency. That is, 70% of the individuals will be routed based on Merit to the

next module, and 30% will be routed randomly to the modules that do not match their proficiency. The third condition, random routing, assigns test-takers based on a coin flip, any examinee will either be correctly or incorrectly routed to the next module. The latter most closely reflects operational procedures in TIMSS, where rotated booklet designs are used (Martin et al., 2013). Previous studies have defined random routing as a comparable baseline to a non-adaptive design (e.g., Rutkowski et al., 2022; Svetina et al., 2019).

Data generation and analysis

The proficiency distribution was generated as a normal distribution with the respective country's mean and standard deviation (see Table 1). Proficiency was generated from random draws that varied across replications but remained equal across conditions. The generation of item parameters follows TIMSS 2015 difficulty and discrimination item parameters for eight-grade mathematics (see Table 2). To simulate students' responses in our MST design, we used the *mstR*[1] (v.1.2., Magis, 2018) package in *R*. While the MST 1–2–3 design was used, the routing probabilities varied, and DIF was induced to random items within the modules following the DIF conditions. All DIF magnitudes, amounts per module, and locations were only introduced to items for either low, medium, or high performing countries. Routing probabilities were used across all DIF and no DIF conditions. Three baseline conditions where all items are DIF-free [3 routing probabilities] and 243 DIF conditions [3 DIF magnitudes × 3 DIF amounts × 3 DIF locations × 3 DIF countries × 3 routing probabilities] sum a total of 246 conditions studied. All DIF conditions were compared to their baseline condition of the same routing probability, with 100 replications performed for each condition.

Once the student responses were generated for each condition, international item parameters for the pooled sample (n = 36,000) were estimated using a multigroup 2-PL IRT model with equality constraints on item parameters across groups with *TAM* (v.3.1., Robitzsch et al., 2018). The difficulty and discrimination item parameter estimates were then transformed to the original scale via a mean-sigma adjustment (Kolen & Brennan, 2004). The re-scaled estimated international item parameters were considered fixed and used to set the scale for the country proficiency parameter estimation, which was done through a latent regression model. The latent mean and variance of the first group were fixed to a standard normal distribution ~ $N(0,1)$ to set the scale. Based on the posterior achievement distribution, five plausible values for mathematics proficiency were drawn for each examinee. Following TIMSS 2015 scaling methodology (Martin et al., 2016) and general procedures for analyzing multiple imputed datasets (Rubin, 1987), descriptive statistics were obtained for each country. Therefore, for each population, a statistic will be equal to $\theta = \frac{1}{M}\sum_{i=1}^{M}\theta_i$, where M is the number of plausible values, and θ_i is the mean (or any other statistic of interest) computed for the ith plausible value.

Evaluation criteria

The impact of DIF on routing was evaluated through module selection accuracy, item parameter, and final proficiency parameter recovery.

[1] Magis modified the *mstR* package to allow for the probabilistic routing element.

Table 2 Item difficulty and discrimination parameters used for data generation

	Stage						
	Module	Core	1		2		
			Easy	Difficult	Easy	Moderate	Difficult
Difficulty parameter	i1	0.318	**0.533**	**0.533**	**0.260**	**0.260**	**0.260**
	i2	0.742	**0.253**	**0.253**	**0.818**	**0.818**	**0.818**
	i3	0.635	**0.266**	**0.266**	**0.256**	**0.256**	**0.256**
	i4	0.247	− 0.052	0.977	− 0.079	0.260	2.163
	i5	0.652	0.048	1.000	− 0.075	0.818	1.472
	i6	0.684	0.013	0.900	− 0.259	0.256	1.517
	i7	0.505	0.112	1.037	− 0.105	0.488	1.313
	i8	0.752	0.190	1.127	− 0.140	0.704	1.450
	i9	0.613	− 0.030	0.887	− 0.251	0.674	1.582
	i10	0.759	0.039	0.885	− 0.231	0.759	1.178
	i11	0.453	0.164	1.050	− 0.174	0.646	1.397
	i12	0.646	0.154	1.105	− 0.224	0.779	1.498
Discrimination parameter	i1	1.473	**0.910**	**0.910**	**1.045**	**1.045**	**1.045**
	i2	1.976	**0.683**	**0.683**	**1.030**	**1.030**	**1.030**
	i3	1.316	**0.705**	**0.705**	**1.143**	**1.143**	**1.143**
	i4	1.126	1.459	1.754	0.692	1.045	0.580
	i5	1.344	0.659	0.874	0.915	1.030	1.890
	i6	1.152	1.098	1.553	1.226	1.143	1.148
	i7	0.855	1.333	1.530	0.681	0.765	1.313
	i8	1.545	1.248	1.326	1.306	1.599	1.237
	i9	1.204	1.352	1.650	1.186	1.764	1.398
	i10	1.399	0.513	1.388	0.821	1.399	1.058
	i11	1.158	1.376	1.147	1.136	1.676	1.577
	i12	1.676	1.389	2.166	0.912	1.136	1.706

Note: Under Stages 1 and 2, the first three items are trend items and therefore are contained in each module of a respective stage. Bold parameters represent trend items that are repeated in each module within a stage.

Module selection accuracy

To inform our results in terms of the effects of DIF in MST routing, we used the proportion of routing decisions that matched between the DIF and no DIF conditions. For the 1–2–3 MST design proposed, routing can occur only after Core and Stage 1 modules. Results that equal one indicate DIF does not affect routing, as the DIF and no DIF conditions match perfectly across all students in a country. Otherwise, differences between the DIF and no DIF conditions exist. Each condition was averaged across examinees within a country and across replications.

Item parameter and proficiency recovery

We used two properties of estimators to evaluate parameter estimates: bias and root mean square error (RMSE). For item parameters (difficulty and discrimination), we obtained bias by obtaining the difference of the estimates to DIF-free conditions, the original international estimates used for data generation. For proficiency estimates, bias is the difference between estimates and true proficiency values for each country. We

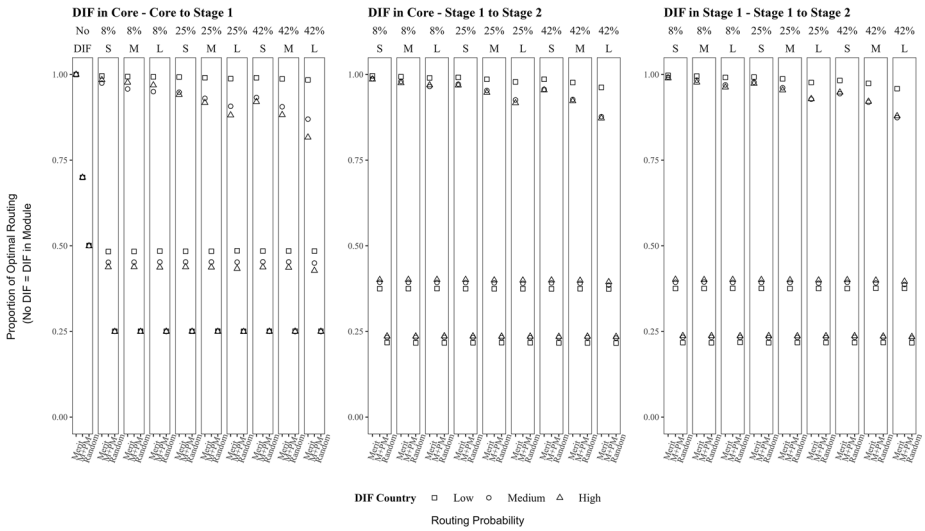

Fig. 2 Proportion of the accuracy in routing to the next module

generally expect a larger bias and RMSE under DIF conditions compared to conditions with no DIF.

Results

Results are reported as follows: Section I presents results for the accuracy of module routing, while Section II presents results for item parameters and proficiency estimation recovery.

Section I: Module Routing Accuracy

In evaluating the effect of DIF on MST routing decisions, we expected to observe lower accuracy in the decision of routing from one module to the next given that a previous module was DIF contaminated. Results related to the module routing accuracy are presented in Fig. 2, which shows three panels, each of which indicates the DIF location and the routing accuracy in the decision made to route towards the next module in Stage 1 and Stage 2. Within those panels, nine inner columns showed the percentage out of 12 items and the magnitude of DIF items per module. The module accuracy selection is on the Y-axis, which ranges from 0.32 to 1.00. The X-axis represents the routing probabilities. To evaluate differences of DIF countries, the shapes represent DIF in low (square), medium (circle), and high-performing (triangle) countries. The baseline condition for each of the three different types of routing was equal to one (see inner column at the left, No DIF). The interpretation of Fig. 2 shows the average proportion of test takers who were assigned to the correct modules when we compare test takers within a country with the same generated proficiency level from no DIF and DIF conditions. In other words, a proportion of 0.32 indicates that on average only 32% of test takers with the same proficiency level in a low performing country with DIF were routed equally in the presence of DIF as they would have been in the absence of DIF.

The impact of DIF for Merit routing showed worse routing accuracy when higher DIF magnitudes and a higher number of DIF items were present in medium and high performing countries. In general, interactions of higher magnitudes of DIF and a greater number of DIF-contaminated items in a module seemed to affect the routing accuracy of Merit routing. To evaluate the impact of DIF in Merit routing, consider the column at the far right of the first panel of Fig. 2 (42%/Large DIF condition), which reported the lowest routing accuracy results; 0.82 for DIF in high, 0.87 for DIF in medium, and 0.98 for DIF in low performing countries. When we evaluate the same far right column (42%/Large DIF condition) for the second panel, we observed that the routing accuracy results slightly improve for DIF in high (0.87) and medium (0.88) performing countries and worsen for DIF in low (0.96) performing countries. Similar average routing accuracy results were observed for the same 42%/Large DIF condition on the third panel. Our findings pointed to a more substantial impact in routing when DIF occurred in the Core. Consider a situation where we extend the accuracy scale to thousands, closer to the number of participants in ILSAs. In that case, we observed that in the worst-case scenario, 19%—or about 190 students for every thousand—would be misrouted in the presence of large DIF magnitudes and more items with DIF per module when routed based on their Merit.

The impact of DIF in M + PM and Random routing showed, on average, similar routing accuracy results regardless of the number of items with DIF in a module or the magnitude of DIF. The accuracy results for Random routing were similar regardless of the country with DIF but varied based on where in the MST design DIF appeared. When DIF was present in the Core module, the routing from the Core to Stage 1 showed, on average, the accuracy was 0.50 for DIF in high, medium, and low performing countries. When routing from Stage 1 to Stage 2, the average accuracy for DIF in low was 0.35, 0.36 for DIF in medium, and 0.37 for DIF in high performing countries. When there was DIF on the modules of Stage 1, the routing between Stage 1 and Stage 2 was 0.36 for all countries that presented DIF. The accuracy results for M + PM routing were slightly higher than for Random routing but considerably lower than Merit routing. When DIF was present in the Core module, the routing accuracy from Core to Stage 1 was 0.58 for DIF in low, 0.56 for DIF in medium, and 0.54 for DIF in high performing countries; the routing accuracy from Stage 1 to Stage 2 was 0.45 for DIF in low, 0.45 for DIF in medium, and 0.46 for DIF in high performing countries.

Figure 3 illustrates the effects of DIF on the proportion of non-optimal routing. In this study's 1–2-3 MST design, two non-optimal paths existed: Core-High-Low and Core-Low–High. Non-optimal paths could only happen on the routing from Stage 1 to Stage 2, as seen by the dashed lines in Fig. 1. Therefore, the results of Fig. 3 show the carryover effects from the DIF on the Core Module and the effects of DIF on Stage 1 in the transition from Stage 1 to Stage 2. The non-optimal routing happened when test-takers in the absence of DIF were routed to any of the four optimal paths (e.g., Core-Low-Low, Core-Low-Med, Core-High-High, Core-High-Med), but in the presence of DIF were routed to a non-optimal path instead. The four panels in Fig. 3 can be interpreted similarly to Fig. 2, except that Fig. 3 illustrates the average proportion of non-optimal routing.

Results of non-optimal routing were equal to zero across all conditions when the MST used Merit routing. As expected, the algorithm only routes examinees based on their

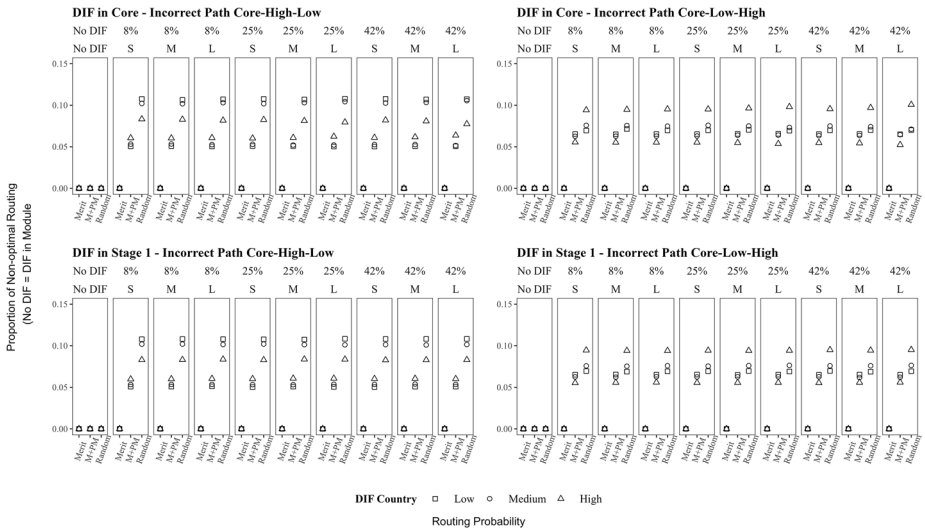

Fig. 3 Proportion of the non-optimal paths selected given DIF

Merit through optimal paths only. However, M + PM and Random routing illustrated how countries with DIF misrouted their test-takers due to the presence of DIF. For example, M + PM routing and DIF in the Core had test-takers from countries with DIF misrouted to the Core-High-Low path (see upper left panel) by 0.06 for DIF in high, 0.05 for DIF in medium, and 0.05 for DIF in low performing countries. Under the same condition, test-takers were misrouted into the Core-Low–High path (see upper right panel), in high proportions for DIF in medium (0.06) and low (0.07) performing countries and lower proportions for DIF in high (0.05) performing countries. The results' patterns were similar for DIF Stage 1 modules (bottom panels) and across DIF amounts; however, the proportions were barely smaller. Slight variations at the third decimal place were noted for DIF magnitudes but not an evident pattern for DIF amounts. Random non-optimal routing results were worse than M + PM, and patterns for DIF countries were the opposite to the results from M + PM. For example, misrouting to the Core-High-Low path when the MST used Random routing and there was DIF in the Core module (see left upper panel) showed more non-optimal paths when DIF was present in low (0.11) performing countries, followed by DIF in medium (0.10) and high (0.08) performing countries. Under the same condition, the Core-Low–High non-optimal path when MST used Random routing selected on average, a higher proportion under the presence of DIF in high (0.10), than in medium (0.07) and low (0.07) performing countries.

Section II: Parameter Recovery

Item difficulty and discrimination recovery

The estimation of difficulty and discrimination parameters was evaluated using a scatterplot between the estimated and generated parameters averaged over 100 replications in each condition for each of the 72 items; results are found in Figs. 4 and 5. In Figs. 4 and 5, three panels show the differences when DIF is present in low, medium, and high performing countries. The columns in Figs. 4 and 5 represent the DIF magnitude and

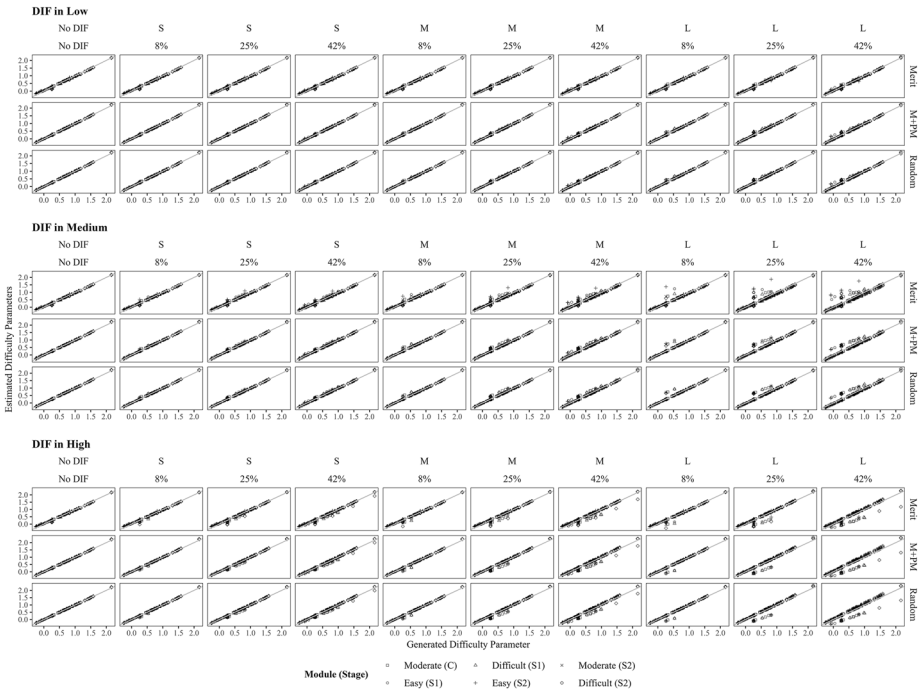

Fig. 4 Difficulty estimates and generating parameter's correlation for all items across DIF conditions

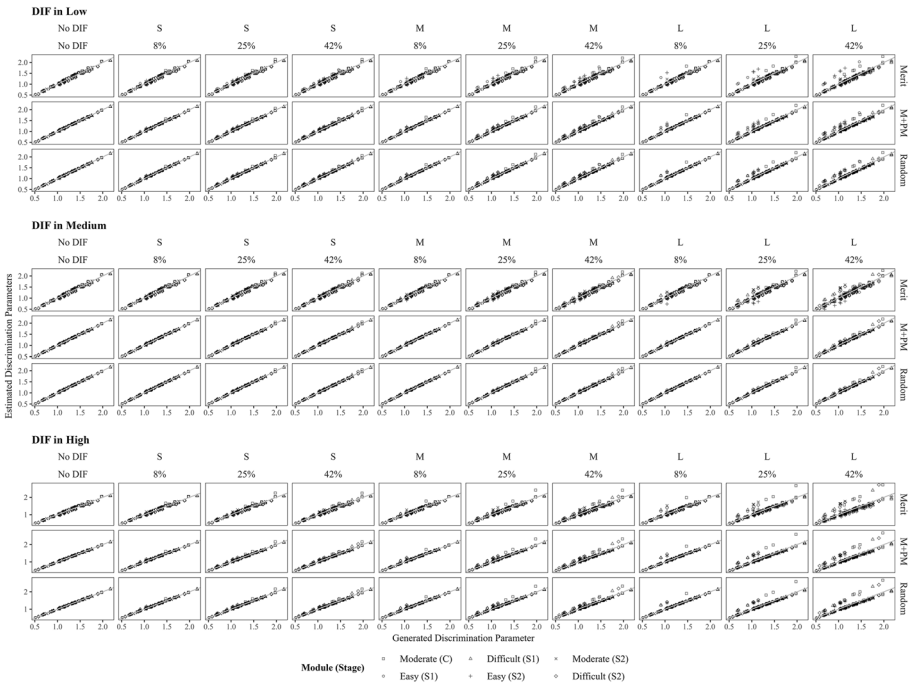

Fig. 5 Discrimination estimates and generating parameter's correlation for all items across DIF conditions

percentage, the rows describe the three routing probabilities, and the shapes (i.e., circle, triangle, square, plus sign) differentiate which module the item belongs. The *y*-axis

shows the estimated parameter, while the x-axis shows the generated parameter. Generated parameters were free of DIF. The diagonal line serves as an indicator of zero bias.

Item difficulty estimates from the multigroup models (where the estimates are equal across countries) showed excellent recovery when items were DIF-free, independently of the routing mechanism used (Fig. 4). The shapes on top of the diagonal line of Fig. 4 indicated that estimated and generated parameters were equal and therefore showed a perfect fit. On the last column to the right, a perfect fit was observed for DIF-free items. A perfect fit was also observed for most items in the rest of the conditions when items did not have any DIF (Fig. 4). This behavior happened for each routing probability with no difference, indicating that item parameter recovery, including a misrouting probability, did not affect the recovery of item difficulty.

Those items that included DIF are away from the diagonal line (Fig. 4). Bias in item difficulty for items with DIF uniformly showed that item difficulty was overestimated when DIF was present in low and medium performing countries and underestimated for high performing countries. More specifically, when a third of the countries had small or large DIF items, the multigroup calibration—where the information of all countries is used to estimate item parameters—absorbed the bias depending on which three countries had DIF. As the DIF magnitude increased and the amount of DIF items per module increased, a larger bias was reported for those items. Depending on the DIF country conditions, the results were either underestimated (i.e., DIF in high) or overestimated (i.e., DIF in low and medium performing). In Fig. 4, the differences for DIF magnitude were evidenced by the distance of the items that had DIF from the diagonal line, and DIF percentages were shown in 1, 3, or 5 items of the same shape away from the diagonal line, respectively. A larger bias was reported when medium and high performing countries had DIF. Another interesting finding is that the results across M + PM and Random routing mechanisms were similar. Still, when Merit routing was implemented, the estimated difficulty parameters were in some cases higher than the generated DIF item difficulties, as seen when medium performing countries had DIF.

Item discrimination parameters also showed some impact from DIF in difficulty parameters. Given that no DIF was simulated in the item slopes, there was expected to be no DIF impact. Nevertheless, as DIF in difficulty parameters was of larger magnitudes and amounts in the modules, the estimated discrimination parameters showed biased results. It is relevant to note that when there was DIF in medium performing countries where M + PM or Random routing was used, the impact of DIF in discrimination parameters was minimal. For low and high performing countries, however, the discrimination parameters were consistently overestimated, particularly when M + PM or Random routing was used in the MST design. Using Merit routing in the MST design showed some discrimination parameters being overestimated and some underestimated.

Proficiency recovery

To evaluate the effects of DIF on proficiency recovery, we focused on bias and RMSE indicators. In Figs. 5 and 6, we observe the bias and RMSE in proficiency estimates. These figures give a more detailed depiction of what drives the bias and RMSE in a country's proficiency using shapes (i.e., circle, triangle, square) representing the MST's routing probability. The columns of Figs. 5 and 6 show the DIF conditions (i.e., amount

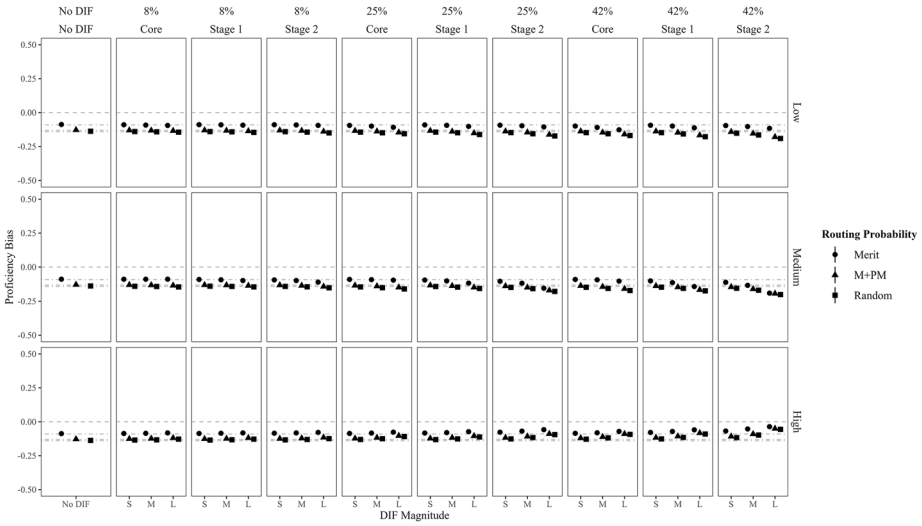

Fig. 6 Bias of proficiency estimates across all conditions

of DIF items per module and DIF items' difficulty in each stage), the x-axis shows DIF magnitudes (i.e., small, medium, and high), with all DIF conditions increasing from left to right. The horizontal blocks in the figures represent high, medium, and low performing countries with DIF, respectively. The y-axis shows the measure of bias and RMSE, respectively. Results were averaged across replications for each condition.

Merit routing showed better proficiency recovery results than M + PM and Random routing. In the absence of DIF, bias results were, on average -0.09 for Merit, -0.13 for M + PM, and − 0.14 for Random routing. In Fig. 6, results that approach the dotted horizontal line at zero had, on average, less bias. Merit routing showed better results across most DIF conditions. The only two exceptions were when the DIF magnitude was large and present in Stage 2 modules. For example, if we take the bias results from the condition where 42% of items had large DIF in Stage 2 modules for medium performing countries (42%/L/Stage2/Medium), we observed bias results of − 0.19 for Merit, − 0.19 for M + PM, and − 0.20 for Random routing. Using the same condition (42%/Large/Stage2/Medium) for high performing countries, we observed bias results of − 0.04 for Merit, − 0.05 for M + PM, and − 0.06 for Random routing. Besides these two conditions, Merit routing showed better proficiency recovery than M + PM and Random routing probabilities.

The proficiency results differed when DIF was present in high performing versus low and medium performing countries. Particularly when there was DIF in large magnitudes and a greater number of DIF items per module, results showed increased bias compared to the no DIF condition for low and medium performing countries but decreased bias results for high performing countries. These results were noticeable when the DIF items were in the later stages of the MST design. If we focus on the farthest right column of Fig. 6 to illustrate the results, where MST using Merit routing and large DIF was present in five items of Stage 2 modules, we found that bias was − 0.12 when DIF was present in low, − 0.19 in medium, and − 0.04 in high performing countries. For the same condition (42%/Large/Stage 2) when M + PM routing was

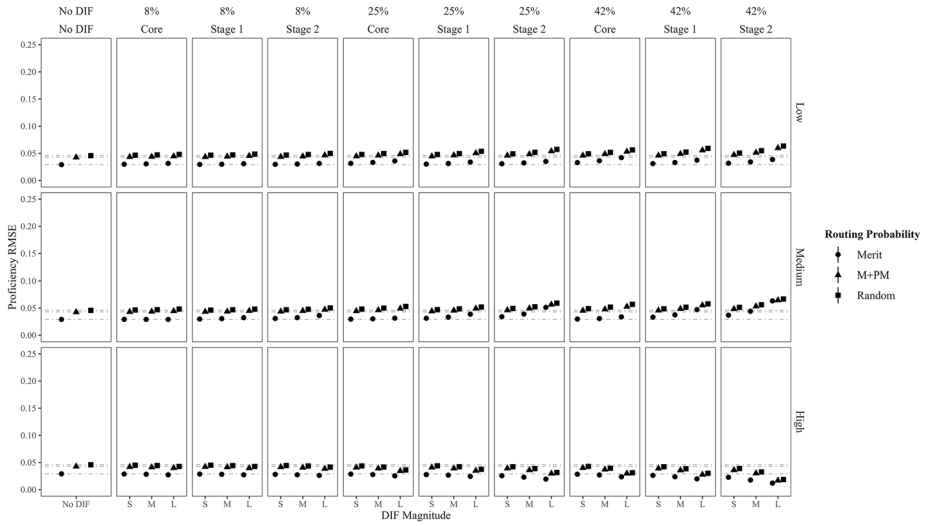

Fig. 7 Root mean square error of proficiency estimates across all conditions

used we observed − 0.18 when DIF was present in low, − 0.19 medium, and − 0.05 high performing countries. Likewise, when Random routing was used, bias was − 0.19 for low, − 0.20 for medium, and -0.06 for high performing countries. There were general similarities between results when DIF is present in low and medium performing countries. Nevertheless, when DIF was present for high performing countries only, the proficiency was recovered better than in the absence of DIF.

DIF location influenced the recovery of the proficiency parameters. When DIF was in the Stage 2 modules, proficiency estimates were more biased and had higher RMSE results when compared to the same DIF conditions in Core and Stage 1. In the Core and Stage 1 modules, the proficiency estimates were essentially unaffected by DIF conditions. Moreover, larger DIF magnitudes and an increasing number of DIF items in a module resulted in more considerable bias and RMSE results. DIF affected proficiency estimates under Merit routing more heavily when DIF items were present in Stage 2 modules than with any other routing mechanism.

Proficiency RMSE results had similar patterns to the bias results. Figure 7 shows the RMSE results averaged across replications. Merit routing reported better proficiency recovery than M + PM and Random routing across all conditions. To illustrate, in the absence of DIF, RMSE for Merit routing was 0.03, 0.04 for M + PM, and 0.05 for Random routing. Worse RMSE results were observed in the presence of DIF in low and medium performing countries. However, a lower RMSE was observed in the presence of DIF in high performing countries, when DIF magnitudes were large, and there were many DIF items per module, especially for Stage 2 modules. For example, when 25% of items in a module had large DIF in Stage 2 modules and Merit routing was used, we observed RMSE results of 0.04 for DIF in low, and 0.05 for DIF in medium, and 0.02 for DIF in high performing countries. Additionally, RMSE results suggested that the presence of DIF items in the Stage 2 modules affected more the recovery results than in the Core or Stage 1.

Discussion

This paper addressed the implications of DIF in an MST design in ILSAs. Regarding routing accuracy, our results showed that DIF propagated to later stages when it occurred earlier in the test. When the Core module had DIF items, the accuracy of routing test-takers from Stage 1 to Stage 2 was as severe as when DIF items were present in Stage 1. This finding implies that subsequent routing errors occur and are driven by early routing errors in the test. Moreover, the accuracy results for Merit routing in the presence of DIF were higher than those for M + PM and Random routing. The large differences in routing accuracy among Merit, M + PM, and Random routing indicated that misrouting decreases the routing accuracy of the same individual across conditions. Moreover, misrouting probabilities counterbalance the effect of DIF amount and magnitude when M + PM and Random routing was used.

Suboptimal routing mechanisms allowed for adaptive paths that did not originally exist in the MST design (i.e., Core-High-Low and Core-Low–High). The creation of these options arises from the 1–2–3 MST design itself. On a 1–2–3 MST design such as the one studied in this paper, suboptimal routing was observed in two different ways, M + PM and Random. These routing mechanisms imply different probabilities from one stage to the next due to the number of modules in the next stage. Suppose there are two modules in the next stage, as from Core to Stage 1, and random routing is implemented; then 50% of the population is assigned appropriately to the next module, and 50% is assigned to the incorrect module, which is similar to a rotated booklet design long implemented by ILSAs. Random routing from Stage 1 to Stage 2 is different because the misrouted population has two options rather than one. In other words, 50% of participants get correctly routed, and 50% are incorrectly routed randomly to either of the two remainder modules, one of which is not within an optimal path. The new adaptive paths could also allow for correcting a previously misrouted individual in the Core to be correctly routed to a module that better fits the test-taker proficiency in Stage 1.

Our findings suggest that the Merit routing mechanism had better recovery of parameter estimates even when DIF was present. However, the exclusive use of Merit routing in ILSAs increases the risk of item overexposure (Svetina et al., 2019) and biased results due to low performing countries consistently routed to easier modules, and high performing countries routed to more difficult ones. As the reader can observe, even Merit routing in the absence of DIF was not equal to zero bias or RMSE. When probabilistic misrouting is implemented, M + PM showed better results than Random routing. Our findings suggest that M + PM routing mechanisms should continue to be implemented in ILSA's MST design based on the provided evidence of proficiency recovery indicators. Importantly, this design offers some protection against DIF. A previous application of M + PM routing in MST was observed in the probability matrix of classification and routing created for PISA 2018 (Yamamoto et al., 2018). Test takers were classified based on their performance in the previous module of the MST design, 90% of students classified as high or low performers were correctly assigned to a module that suited their proficiency in the next stage of the MST design, 10% would be incorrectly assigned; while 50% of students classified as medium performers would get correctly routed and 50% would not. Svetina et al.

(2019) suggested these types of M + PM routing in ILSAs allowed for more control over item exposure. Besides control over item exposure, Yamamoto et al. (2018) supported the argument of content balance. In this paper, we had all countries classified based on a suboptimal routing strategy and based on Merit. The results of this paper support the use of suboptimal routing strategies, particularly when 70% of the population is correctly assigned to the next module, and 30% is incorrectly assigned to the next module.

When DIF was present only for low performing countries in the MST design, we observed two important findings. First, with larger DIF magnitudes and many DIF items in the MST Stage 1 and 2 modules, worse proficiency recovery was observed, specifically when there was DIF in low and medium performing countries; better proficiency recovery was observed when there was DIF in high performing countries. Suboptimal routing, as well as, item difficulty and proficiency distribution mismatch, contribute to these results. Additionally, the worse proficiency recovery for DIF in low or medium countries and the better recovery for DIF in high performing countries could be due in part to having a third of the participant DIF countries (e.g., all low performing countries) getting DIF items (in larger magnitudes and amounts and at later MST stages). The item difficulty estimates for DIF in low and medium performing countries showed more difficult items, and easier items for high performing countries, which in turn reflected in better recovery for DIF in high performing countries. Second, when M + PM routing was used, the MST produced results that minimized the impact of DIF. One explanation could be the misrouting introduced by the M + PM routing, which placed about a third of the population of low performers into the next module based on probability rather than performance, reducing the impact of DIF on routing. DIF in the Core modules showed propagation of accuracy routing into routing in later stages (i.e., Stage 1). Still, the recovery of population proficiency for DIF in the Core remained close to the baseline conditions. Given that ILSAs estimate at the population level, it is possible that these estimates per condition did not capture the propagation of DIF in the Core into later Stages.

Only PISA and PIAAC have implemented an MST design with suboptimal probabilities. However, the most recent administrations of TIMSS and PIRLS have moved into group adaptive testing design (GAT; Mullis & Martin, 2019). The adaptation procedure in GAT follows the logic of longitudinal MST by assigning student groups to a test that better suits their proficiency based on the results of a previous administration. For example, PIRLS 2021 used the estimates of country proficiency from a previous assessment, PIRLS' 2016 results, to inform the routing for country populations. Then easier or more difficult booklets were assigned in varying proportions to populations (70/30 for high and low performing countries and 50/50 for medium performing countries). This study's proposed design could still be implemented in IEA's ILSA studies within one administration and in combination with GAT. Furthermore, the implications found from this study could serve inform future MST design practices.

One of the most critical limitations in the use of adaptive designs is the lack of a sufficient number of items at the end of the proficiency distribution or, more specifically, a mismatch of item difficulty and the proficiency distribution for low-performing countries. The generated difficulty parameters only extended from − 0.26 to 2.16,

while proficiency distribution for the lowest country had a mean and standard deviation of − 1.32 and 0.86, respectively. Almost 90% of students in the lowest performing country will fail to respond correctly to most items in the Core module and will get an easier module based on their Merit, which in turn results in a more accurate estimate of their proficiency. However, given that questions in the easier module are still too difficult for these students, there is little benefit that could be gained from adaptive testing, because we are unable to get enough test information to route the students in the correct direction. The lack of sufficiently easy and difficult items in TIMSS design, as reflected in our results, shows that to fully take advantage of adaptive testing in ILSAs, extending the pool of items with difficulties at the ends of the proficiency continuum will be necessary for low and high performing countries. While not in the context of adaptive testing, Rutkowski et al. (2019) warned the mismatch of item difficulty and proficiency distributions could result in either ceiling or flooring effects. We recommend that test designers include a larger pool of items with difficulties that match the ends of the continuum to fully benefit from the use of adaptive testing in ILSAs.

This study aimed to understand the effects of DIF in an MST design in the context of ILSAs. Based on the results of this simulation study, we found that even when more items with large DIF were present, misrouting inaccuracy rates were small and the proficiency recovery parameters were less affected when the MST design followed a suboptimal routing mechanism, particularly for low and high performing countries. This study only looked at uniform DIF in a 2-PL IRT model; future research on non-uniform DIF and the use of polytomous items could be important, particularly for the estimation of the background questionnaire constructs in ILSAs. Moreover, a one-panel design was implemented in this simulation design to avoid confounding findings from DIF effects in routing. Future research could address a multiple-panel design like the ones used in PISA's MST implementation (Yamamoto et al., 2018) to see if these results remain the same. The mismatch of the population distribution and the item difficulty was a limitation of this analysis, but a portrait of TIMSS 2015 Mathematics results. Even though the selected TIMSS 2015 items do not encompass the whole TIMSS 2015 Mathematics pool, the sample of items still represents that pool of items and the proportions of items with the difficulties studied. Future research in this area should analyze the potential of an adaptive design with suboptimal routing when the mismatch of proficiency distributions and item parameters does not exist, particularly affecting the extremes of the proficiency continuum.

Acknowledgements
We wish to thank Yian-Ling Liaw and Kondwani Mughogho for their observations and feedback in the research meetings when sharing initial iterations of this project. Also, we wish to thank Jeff Yoder for his support on accelerating the simulation code and assistance on language checking.

Author contributions
MVM prepared the simulation study as well as the manuscript. LR provided close guidance concerning the analysis, commented on the several iterations of the manuscript, read, and approved the final manuscript. DSV provided guidance concerning the analysis, read, provided detailed comments, and approved the final manuscript. DR read and provided substantive comments to the manuscript and approved its final version. All authors read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
This is a simulation study, proficiency and item parameters were generated according to TIMSS 2015 mathematics results for eight-graders found online at https://timssandpirls.bc.edu/timss2015/international-database/, and selected countries only served to illustrate low, medium, and high performers in the simulation. A repository with example R code for the simulation and analysis can be found at: https://github.com/Montse11/Impacts-DIF-in-MST-Routing.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 31 January 2022 Accepted: 24 April 2023
Published online: 15 June 2023

References

Ackerman, P. L. (1992). Predicting individual differences in complex skill acquisition: Dynamics of ability determinants. *Journal of Applied Psychology, 77*(5), 598.
Atar, B., Atalay Kabasakal, K., & Kibrislioglu Uysal, N. (2021). Comparability of TIMSS 2015 mathematics test scores across country subgroups. *The Journal of Experimental Education, 91*, 82–100. https://doi.org/10.1080/00220973.2021.1913978
Betz, N. E., & Weiss, D. J. (1974). Simulation Studies of Two-Stage proficiency Testing. *Research Report Minnesota University Minneapolis Department of Psychology 74–4*. https://files.eric.ed.gov/fulltext/ED103466.pdf
Educational Testing Service. (2016). PISA 2018 integrated design. Princeton, NJ: Author. Retrieved from https://www.oecd.org/pisa/pisaproducts/PISA-2018-INTEGRATEDDESIGN.pdf
Ercikan, K. (1998). Translation effects in international assessments. *International Journal of Educational Research, 29*(6), 543–553. https://doi.org/10.1016/S0883-0355(98)00047-0
Gierl, M. J., Lai, H., & Li, J. (2013). Identifying differential item functioning in multi-stage computer adaptive testing. *Educational Research and Evaluation, 19*(2–3), 188–203.
Grisay, A., & Monseur, C. (2007). Measuring the equivalence of item difficulty in the various versions of an international test. *Studies in Educational Evaluation, 33*(1), 69–86. https://doi.org/10.1016/j.stueduc.2007.01.006
Hambleton, R. K., & Swaminathan, H. (1985). A look at psychometrics in The Netherlands.
Hanson, B. A. (1998). Uniform DIF and DIF defined by differences in item response functions. *Journal of Educational and Behavioral Statistics, 23*(3), 244–253. https://doi.org/10.2307/1165247
Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning. Psychology Press.*
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. https://doi.org/10.1207/s15324818ame1903_3
Joo, S., Valdivia, M., Rutkowski, L., & Svetina, D. (in press) Alternatives to weighted item fit statistics for establishing measurement equivalence in many groups. *Journal of Educational and Behavioral Statistics.*
Kara, B. E., & Doğan, N. (2022). The effect of ratio of items indicating differential item functioning on computer adaptive and multi-stage tests. *International Journal of Assessment Tools in Education, 9*(3), 682–696.
Kim, H., & Plake, B. S. (1993). 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, United States*. https://files.eric.ed.gov/fulltext/ED357041.pdf
Kirsch, I., & Lennon, M. L. (2017). PIAAC: A new design for a new era. *Large-Scale Assessments in Education, 5*(1), 1–22.
Kirsch, I., Yamamoto, K., & Khorramdel, L. (2020). Design and key features of the PIAAC Survey of Adults. *Large-scale cognitive assessment* (pp. 7–26). Cham: Springer.
Kolen, M. J., & Brennan, R. L. (2004). Test equating, scaling, and linking: Methods and Practices. *Springer Science & Business Media*. https://doi.org/10.1007/978-1-4939-0317-7
Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates, Inc.
Luo, X., & Kim, D. (2018). A top-down approach to designing the computerized adaptive multistage test: Top-down multistage. *Journal of Educational Measurement, 55*(2), 243–263. https://doi.org/10.1111/jedm.12174
Martin, M. O., Mullis, I. V. S., & Foy, P. (2013). TIMSS 2015 assessment design. In Mullis, I. V. S., & Martin, M. O. (Eds.), *TIMSS 2015 assessment frameworks* (pp. 85–99). TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College. http://timss.bc.edu/publications/timss/2015-methods/chapter-4.html
Martin, M. O., Mullis, I. V. S., & Hooper, M. (2016). *Methods and procedures in TIMSS 2015*. TIMSS & PIRLS International Study Center, Boston College. http://timssandpirls.bc.edu/publications/timss/2015-methods.html
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: Springer. https://doi.org/10.1007/978-0-387-85461-8_9
Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525–543.
Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*(4), 297–334.
Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*(2), 177–196.
Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Chapter 3: Scaling procedures in NAEP. *Journal of Educational Statistics, 17*(2), 131–154. https://doi.org/10.2307/1165166
Mullis, I. V. S., Martin, M. O., Goh, S., & Cotter, K. (Eds.) (2016). *TIMSS 2015 Encyclopedia: Education Policy and Curriculum in Mathematics and Science.* Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/timss2015/encyclopedia/
Mullis, I. V. S., & Martin, M. O. (Eds.). (2019). *PIRLS 2021 Assessment Frameworks*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/pirls2021/frameworks/
Nandakumar, R., & Roussos, L. (2004). Evaluation of the CATSIB DIF procedure in a pretest setting. *Journal of Educational and Behavioral Statistics, 29*(2), 177–199. https://doi.org/10.3102/10769986029002177
Organization for Economic Cooperation and Development (OECD; 2013). Technical report of the Survey of Adult Skills (PIAAC). https://www.oecd.org/skills/piaac/_Technical%20Report_17OCT13.pdf
Organization for Economic Cooperation and Development (OECD) (2019). *PISA 2018 Technical Report*. PISA, OECD Publishing, Paris. https://www.oecd.org/pisa/data/pisa2018technicalreport/
Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling, 53*(3), 315–333.
Pohl, S. (2013). Longitudinal multistage testing. *Journal of Educational Measurement, 50*(4), 447–468.
Robitzsch, A., Kiefer, T., & Wu, M. (2018). TAM: Test analysis modules. R package version 2.10-24.
Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581–592.
Rubin, D. B. (1987). The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm. *Journal of the American Statistical Association*, *82*(398), 543–546.
Rutkowski, D., & Rutkowski, L. (2013). Measuring socioeconomic background in PISA: One size might not fit all. *Research in Comparative and International Education, 8*(3), 259–278. https://doi.org/10.2304/rcie.2013.8.3.259
Rutkowski, D., & Rutkowski, L. (2021). Running the wrong race? The case of PISA for development. *Comparative Education Review, 65*(1), 147–165.
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.
Rutkowski, L., Liaw, Y. L., Svetina, D., & Rutkowski, D. (2022). Multistage testing in heterogeneous populations: Some design and implementation considerations. *Applied Psychological Measurement, 46*(6), 494–508.
Rutkowski, L., Rutkowski, D., & Liaw, Y.-L. (2019). The existence and impact of floor effects for low-performing PISA participants. *Assessment in Education: Principles, Policy & Practice, 26*, 643. https://doi.org/10.1080/0969594X.2019.1577219
Rutkowski, L., Rutkowski, D., & Zhou, Y. (2016). Item calibration samples and the stability of proficiency estimates and system rankings: Another look at the PISA model. *International Journal of Testing, 16*(1), 1–20. https://doi.org/10.1080/15305058.2015.1036163
Svetina, D., Liaw, Y.-L., Rutkowski, L., & Rutkowski, D. (2019). Routing strategies and optimizing design for multistage testing in international large-scale assessments: Routing strategies and optimizing design for multistage testing. *Journal of Educational Measurement, 56*(1), 192–213. https://doi.org/10.1111/jedm.12206
Svetina, D., & Rutkowski, L. (2014). Detecting differential item functioning using generalized logistic regression in the context of large-scale assessments. *Large-Scale Assessments in Education, 2*(4), 1–17. https://doi.org/10.1186/s40536-014-0004-5
Verschoor, A., & Eggen, T. (2014). Optimizing the test assembly and routing for multistage testing. In Duanli, Y., von Davier, A. A., & Lewis, C. (Eds.), *Computerized Multistage Testing: Theory and Applications*. CRC Press/Taylor & Francis Group vey of Adult Skills (PIAAC), Ch. 17 (pp. 406–438).
von Davier, M., & Sinharay, S. (2014). Analytics in international large-scale assessments: Item response theory and population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 155–174). CRC Press.
Wainer, H., Kaplan, B., & Lewis, C. (1992). A comparison of the performance of simulated hierarchical and linear testlets. *Journal of Educational Measurement, 29*(3), 243–251.
Weissman, A. (2016). IRT-based multistage testing. *Computerized multistage testing* (pp. 191–206). Chapman and Hall/CRC.
World Bank. (2021) Country Data. Retrieved from https://data.worldbank.org/country.
Yamamoto, K., Shin, H. J., & Khorramdel, L. (2018). Multistage adaptive testing design in international large-scale assessments. *Educational Measurement: Issues and Practice, 37*(4), 16–27. https://doi.org/10.1111/emip.12226
Yamamoto, K., Shin, H. & Khorramdel, L. (2019). Introduction of multistage adaptive testing design in PISA 2018. *OECD Education Working Papers*, No. 209, OECD Publishing, Paris. https://doi.org/10.1787/b9435d4b-en.
Yan, D., Von Davier, A. A., & Lewis, C. (Eds.). (2016). *Computerized multistage testing: Theory and applications.* CRC Press.
Yin, L. & Foy, P. (2021). TIMSS 2023 Assessment Design. In I.V.S. Mullis, M.O. Martin, & M. von Davier (Eds.), TIMSS 2023 Assessment Frameworks. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2023
Zwick, R. (2010). The investigation of differential item functioning in adaptive tests. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing.* Springer.
Zwick, R., & Bridgeman, B. (2014). Evaluating validity, fairness, and differential item functioning in multistage testing. In D. Yan, A. A. von-Davier, & C. Lewis (Eds.), *Computerized multistage testing.* CRC Press/Taylor and Francis Group.
Zwick, R., & Bridgeman, B. (2016). Evaluating validity, fairness, and differential item functioning in multistage testing. *Computerized multistage testing* (pp. 309–322). USA: Chapman and Hall/CRC.
Zwick, R., & Thayer, D. T. (2002). Application of an empirical bayes enhancement of Mantel-Haenszel differential item functioning analysis to a computerized adaptive test. *Applied Psychological Measurement, 26*(1), 57–76. https://doi.org/10.1177/0146621602026001004

Zwick, R., Thayer, D. T., & Lewis, C. (1997). An investigation of the validity of an empirical Bayes approach to Mantel-Haenszel DIF analysis. *ETS Research Report Series, 1997*(2), i–67. https://doi.org/10.1002/j.2333-8504.1997.tb01742.x
Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement, 36*(1), 1–28. https://doi.org/10.1111/j.1745-3984.1999.tb00543.x
Zwick, R., Thayer, D. T., & Lewis, C. (2000). Using loss functions for DIF detection: An empirical Bayes approach. *Journal of Educational and Behavioral Statistics, 25*(2), 225–247. https://doi.org/10.3102/10769986025002225
Zwick, R., Thayer, D. T., & Wingersky, M. (1994). A simulation study of methods for assessing differential item functioning in computerized adaptive tests. *Applied Psychological Measurement, 18*(2), 121–140. https://doi.org/10.1177/014662169401800203

Publisher's Note