International Journal of Turkish Educational Studies,
2022, 10 (19), 343-354
https://doi.org/10.46778/goputeb.1123902

Effect of Routing Methods on the Performance of Multi-Stage Tests*

Başak Erdem Kara[a]

[a] *Assist.Prof.Dr, Anadolu University, ORCID: 0000-0003-3066-2892*

ABSTRACT

In recent decades, thanks to the great advances and growing opportunities in the technology world, computer-based testing has become a popular alternative to the traditional fixed-item paper-pencil tests. Specifically, multi-stage tests (MST) which is a kind of CBT and an algorithm-based approach become a viable alternative to traditional fixed-item tests with important measurement advantages they provided. This study aimed to examine the effect of different routing rules and scoring methods under different ability distributions in MSTs. For this purpose, three different routing rule, three different ability estimation methods, and two different ability distributions were manipulated in a simulation design. Although there was no clear best method in the studied conditions, it was seen that the Kullback-Leibler was the most efficient routing method and worked best with the EAP scoring method in most of the conditions. Furthermore, EAP and BM provided higher measurement efficiency than the ML method. Recommendations for using those routing methods were provided and suggestions were made for further research.

Article Type
Research

Article Background
Received:
31.05.2022
Accepted:
29.06.2022

Key Words
Multi-Stage Testing,
Routing,
Ability Estimation

To cite this article: Erdem Kara, B. (2022). Effect of routing methods on the performance of multi-stage tests. *International Journal of Turkish Educational Sciences, 10 (*19), 343-354.

Corresponding Author: Başak Erdem Kara, e-mail: basakerdem@anadolu.edu.tr

* This study was a simulation study based on PISA data that can be freely downloaded by everyone from OECD website. Therefore, ethics committee approval was not required.

Uluslararası Türk Eğitim Bilimleri Dergisi,
2022, 10 (19), 343-354
https://doi.org/10.46778/goputeb.1123902

Yönlendirme Yöntemlerinin Çok Aşamalı Testler Üzerindeki Etkisi*

Başak Erdem Kara[a]

[a] *Dr.Öğr.Üyesi, Anadolu Üniversitesi, ORCID: 0000-0003-3066-2892*

ÖZET

Özellikle son yıllarda, teknoloji dünyasındaki gelişmeler ve artan olanaklarla birlikte, bilgisayara dayalı testlerin popülerliği artmış ve bu testler geleneksel kâğıt-kalem testlerin yerine uygulanabilir bir alternatif halini almıştır. Bilgisayara dayalı testlerin bir türü olan ve algoritmaya dayalı bir yaklaşım olan Çok Aşamalı Testler de sağladıkları önemli avantajlarla birlikte kağıt-kalem testlerinin önemli bir alternatifi haline gelmiştir. Bu çalışmanın amacı, Çok Aşamalı Testlerde yönlendirme yöntemlerinin test performansına etkisinin farklı koşullar altında incelenmesidir. Bu amaçla bir simülasyon çalışması tasarlanmış, üç farklı yönlendirme kuralı, üç farklı yetenek kestirim yöntemi ve iki farklı yetenek dağılımı manipüle edilmiştir. Analizler sonucunda hem normal hem de uniform dağılım için, birçok koşulda Kullback-Leibler'in en etkili yönlendirme yöntemi olduğu ve koşulların çoğunda EAP puanlama yöntemiyle en iyi şekilde çalıştığı görülmüştür. Ayrıca, EAP ve BM yetenek kestirim yöntemleri, ML yönteminden daha yüksek ölçüm hassasiyeti sağlamıştır. En düşük ölçüm hassasiyeti ise, tesadüfi yönlendirme yönteminde elde edilmiştir. Bu yönlendirme yöntemlerinin kullanımına ve ileriki araştırmalara yönelik bazı önerilerde bulunulmuştur.

MAKALE BİLGİSİ

Makale Türü
Araştırma

Makale Geçmişi
Gönderim tarihi:
31.05.2022
Kabul tarihi:
29.06.2022

Anahtar Kelimeler
Çok Aşamalı Testler,
Yönlendirme,
Yetenek Kestirimi

Atıf Bilgisi: Erdem Kara, B. (2022). Yönlendirme yöntemlerinin çok aşamalı testler üzerindeki etkisi. *Uluslararası Türk Eğitim Bilimleri Dergisi, 10* (19), 343-354.

Sorumlu yazar: Başak Erdem Kara, e-posta: basakerdem@anadolu.edu.tr

* Bu çalışma, OECD web sitesinden herkes tarafından indirilebilen PISA verilerine dayalı yapılmış bir simülasyon çalışmasıdır. Bu nedenle, çalışmada etik kurul iznine ihtiyaç duyulmamıştır.

Introduction

Thanks to the great developments in technology and computer worlds in recent decades, computer-based testing (CBT) has become more and more popular (Yan, Lewis & von Davier, 2014). As computers provide opportunity for processing large data sets and detailed psychometric analyses, use of item response theory that had been developed before but was not implementable prior to the computers' existence has become practical. The practical implementation of item response theory on computer adaptive tests have resulted in major changes in the way psychological testing is done (Wainer, 2000).

Computer adaptive testing, (CAT) which is a kind of computer-based testing, is an algorithm-based approach designed to present each next item according to the examinees' estimated proficiency level. Ability estimation is updated after each item and the next item is selected to be compatible with updated ability of examinee (Lord, 1980; Yan et. al., 2014). Computer adaptive tests have gained popularity since they provide more efficient and precise measurement with fewer number of items than linear tests especially for the examinees which take part in the upper and lower end of the ability spectrum (Hendrickson, 2007; Lord, 1980; Wainer, 2000).

Multi-stage test is a special kind of computer adaptive tests in which the adaptation occurs at module (item sets) level instead of item level (Yan et al., 2014). Examinees take item sets which are named as module and their ability is estimated based on their responses on that module. After that, s/he is routed to the next module at the next stage. The route, an examinee follows between stages, is called a path and different combinations of modules, stages and paths are called panels which can be thought as parallel forms in linear tests (Hendrickson, 2007; Svetina, Liaw, Rutkowski & Rutkowski, 2019; Wang, Lin, Chang, & Douglas, 2016).

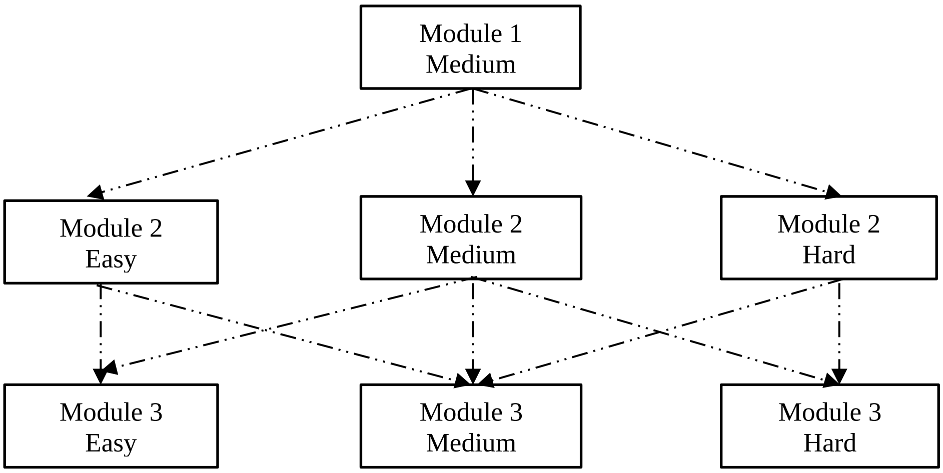

Figure 1. An Example of MST Design

One example of an MST that is a 1-3-3 design 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. In that design, one module at first stage and three modules at second and third stages are involved. Various paths that can be taken by an examinee is shown with arrows. Examinees begin with Stage 1 core module. Depending on the performance on that module, they are directed to the convenient module at Stage 2. After that, performance on previous stages is taken into consideration again and the examinee is routed into one of the Stage 3 module which is appropriate for his/her ability level. As a result of the process, each examinee gets three modules in total each of which from one stage (One module form Stage1, one from Stage 2 and one from Stage 3).

MSTs can be defined as a compromise between CATs and linear tests with features from both design since it combines most of the advantages from CATs and linear tests and minimize their disadvantages (Hendrickson, 2007; Magis, Yan & von Davier, 2017; Yan et. al., 2014). They provide test developers to get more control on test assembly. Each module can be constructed so as to have desired contextual and statistical features. Besides, examinees have a chance to review some items and change previous answers within each module (Hendrickson, 2007; Wainer, 2000; Wang, 2017). Those points make MST usage more common and popular.

With these advantages and growing interest and popularity, MSTs are getting a common usage in real world testing applications (Magis et al., 2017). One of the first international large-scale assessments implementing adaptive testing operationally was the 'The Programme for the International Assessment of Adult Competencies (PIAAC)'. Adaptive test design firstly was used starting in PIAAC 2012 in the form of MST (Yamamoto et al., 2019). In that assessment, it was stated that linear test design provided the same amount of test information with approximately 15-47% more items than the MST design based on the identical item set. Besides, MST was more informative than linear test design through all proficiency ranges (OECD, 2013). Furthermore, PISA which is the largest international large-scale assessment in the world used computer adaptive test designs starting from 2018 cycle. PISA 2018 introduces multi-stage adaptive test design in reading assessment. In that MST designs, individuals did not take fixed, predetermined test booklets. In accordance with the nature of MSTs, they took the assessment as dynamically determined based on their performance on prior stages (Yamamoto et al., 2019). Yamamoto et al. (2019) conducted a simulation study to examine the item response theory model parameter recovery, precision of the person proficiencies and measurement precision that MST design provided. They found that MST design provided an increased precision of 4.5% on average with gains up to 10% at the extreme performance levels with acceptable errors in item parameter estimation.

There are some main elements of an MST application such as item pool, IRT model, panel design, test assembly algorithms, routing methods and scoring methods. All features should be carefully planned and implemented during the process (Wang, 2017). Considering the widespread use and advantages of MST, those elements should be investigated in a more detailed and comprehensive way. Zenisky, Hambleton and Luecht (2010) stated routing methodologies as an important and understudied part of MST and addressed it as the 'next logical direction for research attention'. In each MST application, there should be a routing method controlling the selection of modules, an item pool to form the test and an ability estimation method for interim and final ability estimations. Routing method is the algorithm used to classify test takers to the next stage modules based on their performance. It is important to consider not only the efficiency the MST provided but also the accuracy related with the MST routing decision since it may cause erroneous ability estimations (Sarı, 2016; Yan et al., 2014).

Purpose of Study

Although the importance of routing decision in MSTs are clear, routing methodologies are an understudied part of MST (Zenisky, Hambleton and Luecht, 2010). Studies on routing methodologies on MST are limited to a few number of studies and few number of methods (Kim, Moses & Yoo, 2015; Svetina, Liaw, Rutkowski & Rutkowski, 2019; Weismann, Belov & Armstrong, 2007). The current study aims to investigate which method of routing and scoring works best under different ability distributions. Therefore, the results of the study are likely to contribute to the literature related to routing methods on MSTs. In the context of this study, the following research questions were addressed;

- How does the test performance of MST change with respect to different routing methods (Fisher information, Kullback-Leibler and random) under different ability estimation methods (BM, ML and EAP) and ability parameters' distribution (uniform and normal)?

Method

Within the scope of this research, it was aimed to examine the effect of routing methods on the efficiency of MST applications under different conditions. A Post Hoc simulation study was utilized in order to answer the research question in the study. Simulation study was preferred since it was difficult to meet all conditions at the same time in real data. The simulation process was completed on RStudio environment with mstR package (Magis et. al., 2017).

In the study, several factors were fixed; sample size (n=2000), number of items per design (n=30) and the MST design was set as 1-2-2. Besides 2PL model was used as IRT model. The reason for the choice of MST design and IRT model was that they were the design and model preferred in PISA 2018. The item pool was formed based upon the real item parameters on MST part of PISA 2018. In PISA 2018, there were 245 Reading items in the MST part in total (223 dichotomous and 22 polytomous). For that study, only dichotomously scored items were used to form the item pool, so that the item pool has been formed with the parameters of that 223 items on PISA. Since this was a simulation study based on a freely downloaded data from OECD website, ethics committee approval was not required in the context of this study.

The manipulated factors were routing methods with three levels (MFI, MKL and random), ability estimation method with three levels (ML, BM and EAP), and ability distribution with two levels (uniform and normal distribution).

Routing method. There are mainly two popular kinds of routing methods; number-correct (NC) scoring and IRT based methods (Sarı & Raborn, 2018; Svetina et.al., 2019). In NC method, examinees are routed to the next module based on the number of correct answered items they provided. On the other hand, IRT based methods takes the module information function into consideration while deciding the next module. That module is selected based upon the examinees estimated ability level (Yan et al, 2014). Maximum Fisher Module Information (MFI) and Maximum Module Kullback-Leibler (MKL) are the information-based methods and routes the examinee to the module maximizing the information. The difference between those methods is that they define 'information' in different ways (Weismann et al., 2007). MFI maximizes the module information function and decide on the next module based on that. MKL, on the other hand, maximizes the module Kullback-Leibler information function (weighted by the likelihood function) (Magis et al., 2017). Another method used to route examinees into the next modules is the random selection method which meant that the examinees were routed to the next module randomly. All examinees have the equal chance of being distributed to the any of next module regardless of provisional ability estimate (Svetina et. al., 2019). There were some studies to compare those routing methods in the literature. Kim, Moses, and Yoo (2015) stated that NC scoring was more practical to use since it is a fairly straightforward method and easier to understand while IRT methods are more beneficial psychometrically since they provide more precise estimation. Weismann, Belov and Armstrong (2007) compared three routing methods (Number-Correct Routing and two information-based routing; Maximum Fisher Information, Maximum Mutual Information) in a simulation study. As a result, they suggested that NC routing method is preferable for practical testing purposes and it gave similar classification rates to information-based methods although the information-based methods were able to classify a slightly higher percentage of examinees. Besides, despite of the higher classification accuracy, those methods resulted with the expense of item (over) exposure, particularly in later MST stages. Another research conducted by Svetina, et.al. (2019) investigated whether NC scoring, IRT based methods and random routing were differentially effective at routing students to the correct module and precision of the test in a simulation study. As a result, they found that IRT based methods (both EAP and information based) and random routing method were most successful on theta estimates' recovery.

In the current study, mainly information-based methods and random routing method were the main focus of the study. Three different routing methods were compared in order to decide on the next module: (1) Maximum Fisher Information (MFI) (2) Maximum Kullback-Leibler (MKL) (3) random selection which next module is selected randomly with equal probability for each module.

Ability estimation methods. Three different methods were used in the context of that research: (1) ML (2) BM and (3) EAP. Those were three of the most common and popular methods used in adaptive testing contexts (Magis et. al., 2017).

Maximum likelihood (ML) estimation method maximizes the likelihood function ($L(\theta)$) with respect to θ to get ability estimation.

$$L(\theta) = \prod_{j=1}^{J} \prod_{k=0}^{K_j} P_{jk}(\theta, p_j)^{Y_{jk}}$$

Y_{jk}: Equals one when response category k for item j was chosen and zero otherwise

ML is an asymptotically unbiased and consistent method. However, the problem with the ML method is that it takes infinite values if there is a constant pattern of only correct or only incorrect responses) (Magis et al., 2017).

Bayes model (BM) estimation obtains the θ value by maximizing the posterior distribution (product of the likelihood function and the prior distribution) of ability. The equation that is maximized is the following one. *g(θ)* represents the posterior distribution and *f(θ)* represents for the prior one.

$$g(\theta) = f(\theta)L(\theta) = f(\theta)\prod_{j=1}^{J}\prod_{k=0}^{K_j} P_{jk}(\theta, p_j)^{Y_{jk}}$$

On the other hand, θ can be obtained by computing the expected value instead of maximizing the posterior distribution. That is the case for expected A Posteriori (EAP) method.

$$\hat{\theta} = \frac{\int_{-\infty}^{+\infty} \theta\, g(\theta)\, d\theta}{\int_{-\infty}^{+\infty} g(\theta)\, d\theta}$$

Both BM and EAP methods are based on Bayesian statistics and their accurate estimation are based on the correct selection of prior ability distribution.

Ability distribution. Ability parameters of 2000 examinees was drawn in two different ways (normal and uniform distribution); N(0,1) and U(-3, +3).

Test design. In that study, one MST test design (MST 1-2-2) was investigated under different ability estimation methods, routing methods and ability distribution. All manipulated conditions were fully crossed with each other which resulted in 18 conditions (3 routing methods, 3 ability estimation method and 2 ability distribution) and 30 replications were performed for each condition. About the number of replications that should be used, there is no definite answer since it depends on the purpose and conditions of study (model complexity, number of factors etc.). Harwell et al. (1996) suggested a minimum number of 25 replications in IRT-based research. Considering the practical constraint of time per replication, it was thought that 30 replication is enough in the context of that study.

Detailed information on how MST simulation was conducted is given here. In order to create the environment, mstR package (Magis, Yan, von Davier, 2018) was used and procedures were carried out on the RStudio environment. In the simulation, 18 conditions in total, including three routing method, three ability estimation method and two ability distribution were examined. In 1-2-2 MST design, a single module was used in the first stage, while there were two modules each in the second and third stages. Each individual answered three modules in total, each one from different stages. Items for those modules were randomly selected from item pool so as to take item distribution on modules in PISA into consideration. For instance, Stage 1 items were randomly taken from PISA Stage 1 items. Stage 2_Easy module items were taken from the PISA Stage 2_Easy items etc.

Data Analysis

In order to examine the efficiency of MST; Root Mean Square Error (RMSE), bias, and correlation values between estimated and true ability parameters were used. RMSE, bias and correlation values were calculated for each of the 30 replications and interpretations were made based on the average of those values. Related formulas were presented below.

$$RMSE = \sqrt{\frac{\sum_{j=1}^{N}(\hat{\theta}_j - \theta_j)^2}{N}} \qquad Bias = \frac{\sum_{j=1}^{N}\left|(\hat{\theta}_j - \theta_j)\right|}{N}$$

$\hat{\theta}_j$: Estimated ability parameter θ_j : True ability parameter N : Total number of individuals.

$$\rho_{\hat{\theta}_j,\theta_j} = \frac{cov\,(\hat{\theta}_j,\theta_j)}{\sigma_{\hat{\theta}_j}\sigma_{\theta_j}}$$

$\sigma_{\hat{\theta}_j}$: standard error values of estimated ability parameters σ_{θ_j} : standard error values of true ability parameters

Findings

In this study, the performance of three routing procedures under different ability estimation methods and ability distribution were examined. Findings of MST simulations were presented in detail below.

First of all; RMSE, bias and correlation values when ability distribution is normal is given at Table 1 below.

Table 1. RMSE, Bias and Correlation Values When Ability Distribution is Normal

		Ability Estimation Methods								
		RMSE			Bias			Correlation		
		BM	ML	EAP	BM	ML	EAP	BM	ML	EAP
Routing Methods	MFI	0.398	0.478	0.381	0.003	-0.006	0.000	0.917	0.912	0.924
	MKL	0.383	0.449	0.382	0.003	-0.004	0.001	0.924	0.920	0.925
	Random	0.412	0.530	0.394	0.004	-0.006	0.000	0.911	0.898	0.920

According to the values at Table 1, using 'random' method for routing resulted with the highest RMSE [0.394, 0.530] and lowest correlation [0.898, 0.920] for each of three ability estimation methods. Although bias values are so close to each other, 'random' method has the highest bias [-0.004, -0.006] at most of the condition. It gave the lowest RMSE and highest correlation and bias at ML method and worked best with the EAP method. On the other hand, MKL method had the lowest RMSE and highest correlation [0.920 - 0.925] under most of the conditions. Overall, it was indicated that MKL, which had the lowest RMSE and highest correlation values, had the highest measurement precision. Besides, it worked best with the BM and EAP ability estimation methods. On the other hand, random method that had the highest RMSE and lowest correlation has the lowest measurement precision especially under ML ability estimation method.

Tablo 2 indicates RMSE, bias and correlation values when distribution of ability parameters of individuals was uniform.

Table 2. RMSE, Bias and Correlation Values When Ability Distribution is Uniform

		Ability Estimation Methods								
		RMSE			Bias			Correlation		
		BM	ML	EAP	BM	ML	EAP	BM	ML	EAP
Routing Methods	MFI	0.498	0.613	0.428	-0.001	0.007	0.0003	0.963	0.954	0.971
	MKL	0.461	0.610	0.431	0.000	0.001	0.001	0.968	0.954	0.971
	Random	0.518	0.637	0.459	-0.004	0.003	0.002	0.961	0.951	0.967

Similar to the condition that the ability distribution was normal, random routing method had the highest RMSE and lowest correlation values through all ability estimation methods. On the other hand, MKL had the lowest RMSE [0.431 – 0.610] and highest correlation [0.954 – 0.971] at each of three ability estimation methods. Furthermore, bias values revealed that MKL had lower bias values at most of the conditions regardless of estimation method. Those results indicated that MKL has the best and random method has the worst measurement precision similar to the condition that ability distribution was normal. MKL worked best with the EAP ability estimation. Negative bias values which mean the underestimation of true ability values were only obtained for BM ability estimation method. However those values are very close to zero.

Discussion

In the context of that study, 18 separate MST simulation with different routing methods (MFI, MKL and random), ability estimation methods (ML, BM and EAP) and ability distribution (uniform and normal) were conducted.

One of the main findings was that, MKL routing method had the highest measurement efficiency for both normal and uniform ability distribution. Besides, it worked best with the EAP ability estimation method at most of the conditions. It was expected that the information-based methods worked better than the random routing method since it based on a systematic to route examinees rather than selecting modules randomly. The difference between the MKL and MFI methods is the way that they define the 'information'. KL is defined as "global information" since it measures the discrimination power of Θ_0 and Θ_1 whether they are close together or not. FI is called, on the other hand, "local information" and only measures the item discrimination close to θ_1 (Wang, Chang and Boughton, 2011). The difference between MKL and MKI obtained as a result of the study may be the result of the way that they use to estimate the information.

The results of the current study also indicated that EAP and BM ability estimation methods which both were based on Bayesian statistics worked better than the ML method and that finding is in line with the literature. Haberman & von Davier (2014) also stated that ML estimator caused more problematic consequences on routing. The reason may be that a large number of items are required and it is not defined for the examinees with extreme scores (Haberman & von Davier, 2014). In addition to that, Diao and Reckease (2009) compared the ML and Bayesian methods and indicated that Bayesian methods outperformed the ML method especially for the short length tests. In general, between two Bayesian methods, EAP worked better than the BM method.

In summary, the results of the study provide some guideline for researchers/testing organizations working on adaptive testing implementation. Main finding was that the MKL method worked best at most of the conditions for both uniform and normal ability distributions and it generally worked best with the EAP ability estimation method. In line with these findings, it can be suggested that, MKL could be preferred especially with EAP method regardless of ability distribution. On the other hand, ML ability estimation method and random routing method should be used carefully if they are preferred. As a simulation study, this research has some limitations. The results of that study are limited with the item pool and manipulated conditions. For instance, only 1-2-2 panel design and 30-item test length were used. Further researches can be made by using different MST panel designs, test length, items pools and routing methods. In addition to that, only dichotomous items were used in that study. Test designs including polytomous items can be the subject of further researches. Besides, that is a simulation study although the item parameters were obtained from PISA 2018. So, it cannot be guaranteed that same results would be obtained in an empirical environment.

Author Contributions

All stages of the study were carried out by the author.

Conflict of Interest

The author declares that there is no potential conflict of interest.

References

Diao, Q., & Reckase, M. (2009). Comparison of ability estimation and item selection methods in multidimensional computerized adaptive testing. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*. Retrieved [13.03.2021] from https://www.psych.umn.edu/psylabs/CATCentral

Haberman, S. J. & von Davier, A. A. (2014). Considerations on parameter estimation, scoring, and linking in multistage testing. In D. Yan, A. A. von-Davier & C. Lewis (Eds.), *Computerized multistage testing* (p. 229 – 246). CRC Press; Taylor&Francis Group.

Harwell, M., Stone, C. A., Hsu, T., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, *20*(2), 101–125.

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

Kim, S., Moses, T., & Yoo, H. (2015). A comparison of IRT proficiency estimation methods under adaptive multistage testing. *Journal of Educational Measurement*, 52(1), 70-79.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Mahwah, New Jersey: Routledge

Magis, D., Yan, D. & von-Davier, A. (Eds.). (2017). *Computerized adaptive and multistage testing with R: Using packages catR and mstR*. Springer.

Magis, D., Yan, D. & von Davier, A., A. (2018). Package 'mstR': Procedures to generate patterns under multistage testing. Retrieved from https://cran.microsoft.com/snapshot/2018-09-29/web/packages/mstR/mstR.pdf

OECD (2013). *Technical report of the survey of adult skills (PIAAC)*. OECD Publishing: Paris, France.

Sarı, H. İ. (2016). *Examining content control in adaptive tests: Computerized adaptive testing vs. computerized multistage testing.* Unpublished doctoral dissertation. University of Florida, USA.

Sarı, H. İ., & Raborn, A. (2018). What information works best?: A comparison of routing methods. *Applied Psychological Measurement*, 42(6), 499–515. DOI: 10.1177/0146621617752990

Svetina, D., Liaw, Y. L., Rutkowski L. & Rutkowski, D. (2019). Routing strategies and optimizing design for multistage testing in international large-scale assessments. *Journal of Educational Measurement*, *56*(1), 192-213.

Wainer, H. (2000). Introduction and history. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd Edition), (p. 1–22). Lawrence Erlbaum Associates.

Wang, S., Lin, H., Chang, H. H., Douglas, J. (2016). Hybrid computerized adaptive testing: From group sequential design to fully sequential design. *Journal of Educational Measurement*, 53, 45-62.

Wang, K. (2017). *A fair comparison of the performance of computerized adaptive testing and multistage adaptive testing.* Unpublished doctoral dissertation. Michigan State University, USA.

Weissman, A., Belov, D.I., & Armstrong, R.D. (2007). *Information-based versus number-correct routing in multistage classification tests* (Research Report RR-07–05). Law School Admissions Council.

Yamamoto, K., Shin, H. J. and Khorramdel, L. (2019), *Introduction of multistage adaptive testing design in PISA 2018.*(OECD Education Working Papers, No. 209). OECD Publishing: Paris. DOI: 10.1787/b9435d4b-en.

Yan, D., Lewis, C & von-Davier, A. A. (2014). Multistage test design and scoring with small sample. In D. Yan, A. A. von-Davier & C. Lewis (Eds.), *Computerized multistage testing* (p. 303–324). CRC Press; Taylor&Francis Group.

Zenisky, A., Hambleton, R. K. & Luecht, R. M. (2010). Multistage testing: Issues, design, and research. In W. J. van der Linden & C. A.W. Glas (Eds.), *Elements of adaptive testing* (p. 355-372). Springer.