International Journal of Psychology and Educational Studies, 2020, 7 (1), 27-37

PES

www.ijpes.com

International Journal of Psychology and Educational Studies

Testing Multistage Testing Configurations: Post-Hoc vs. Hybrid Simulations

Halil Ibrahim Sari

Kilis 7 Aralik University

ARTICLE INFO

Article History:
Received 28.09.2019
Received in revised form 12.11.2019
Accepted 19.11.2019
Available online 31.01.2020

ABSTRACT

Due to low cost monte-carlo (MC) simulations have been extensively conducted in the area of educational measurement. However, the results derived from MC studies may not always be generalizable to operational studies. The purpose of this study was to provide a methodological discussion on the other different types of simulation methods, and run post-hoc and hybrid simulations in the multistage testing environment (MST), and to discuss the findings and interpretations derived from each simulation method. The real data collected via paper-pencil administrations from 652 students were used to test different MST design configurations under different test lengths. The levels of test lengths were 24-item, 36-item and 48-item. The five levels of MST designs were 1-2, 1-3, 1-2-2, 1-3-3, 1-4-4 and 1-5-5 design. Both post-hoc and hybrid simulations were run with the same group of students. All analyses were completed in R program. The results indicated that in terms of absolute bias, RMSE and pearson correlations, post-hoc and hybrid simulations generally resulted in comparable outcomes. Regardless of the simulation method, the 1-5-5 MST design was the most recommended design. Another finding was that in all simulations, as the test length increased, the outcomes were better, in general. Advantages and disadvantages of each method, and recommendations for practitioner and limitations for future research were provided in the study.

Keywords:
multistage testing, post hoc, hybrid simulation

1. Overview of Multistage Testing

Multistage testing (MST) is a computer adaptive testing method that became popular following its use in 2011 by the Educational Testing Service (ETS) in the Graduate Record Examination (e.g., r-GRE) in 2011 (Davey & Lee, 2011; Yan, von Davier & Lewis, 2014). The MST is similar to the computerized adaptive testing (CAT) but it tailors a set of items (e.g., 15 items) instead of individual items as the CAT does. In the MST, pre-constructed item sets (i.e., known as module) are placed in the stages, and stages (e.g., division of a test) are placed in the panels (Luecht & Sireci, 2011). In stage one, there is typically one module, called routing module, at the medium difficulty level. The modules in other stages are at the different difficulty levels such as easy, medium and hard. This means that each module provides more information for low, medium and high proficiency test takers. A test developer can arbitrarily vary the number of items in modules, the number of modules in stages, the number of stages in panels, and even the number of panels (Hendrickson, 2007). Increasing the number of panels is essential for test security purposes, especially in high stake tests.

In Figure 1, there is an illustration of MST configuration. In this example, there is one module in stage one, and five modules from easiest to hardest in stage two and three. This configuration is named the 1-5-5 panel

Corresponding author's address: Department of Educational Science, Muallim Rifat College of Education, Kilis 7 Aralik University
Telephone: 0348 814 26 66 ETX: 1677
e-mail: hisari87@gmail.com

http://dx.doi.org/10.17220/ijpes.2020.01.003

design. A test taker starts the exam by taking the routing module placed in stage 1, and based on his/her performance on this starting module, he/she receives, one of the available modules in stage two. The same thing continues until the test taker completes all of the stages. Then, the computer algorithm calculates the final proficiency estimate of the test taker. Based on the previous work, MST may provide more accurate proficiency (e.g., ability level) estimates with higher number of stages in panels and the number of modules in stages (Patsula, 1999; Zenisky, 2004).

The advantages of multistage testing over CAT include allowing item review, item skip and providing greater control to the test developer (Liu, Bridgeman, Gu, Xu, & Kong, 2015). Even though it is disallowed to revise items in the previous modules, one can still go back to the previously administered items within a module. However, the traditional CAT does not give this chance to a test taker. Similarly, the traditional CAT forces examinees to pick one of the response options for an item, but MST gives examinees the right to skip any test item. A well-known advantage of CAT over MST is producing slightly better measurement accuracy in proficiency estimate due to the item level adaptation (Yan et al., 2014).

Even though MST is relatively new compared to the CAT, there are many monte-carlo studies that have been conducted to test many aspects of MST including proposal of different test assembly methods (Luetch, 2000; Luecht & Nungester, 1998), stage and module configuration (Patsula, 1999), content balancing issues (Sari & Huggins-Manley, 2017), and new routing rules (Luetch, 2000; Sari & Raborn, 2018; Thissen & Mislevy 2000). All of these monte-carlo (MC) simulation studies are highly valuable in terms of making a forecast for multiple research questions in the area of psychometrics and finding better solutions for the problems encountered in operational studies (Feinberg & Rubright, 2016). However, MC simulations generally rely on strong assumptions such as no incomplete data, well-fitting with the model and data, no software bugs and ideal optimizations with automated test assembly (ATA) for module building. Unfortunately, these assumptions can be violated in real administrations, and violations may affect the test outcomes. For example, many simulations use Birnbaum's (1968) three parameter (3PL) item response theory (IRT) model, however the 3PL model does not always converge in operational applications of IRT models (de Ayala, 2009). Also, while it is much easier to find a solution for panel and module assembly in simulations, optimization problems can always occur in an operational study. Thus, there is always a possibility that recommendations and interpretations derived from MC simulation studies may not generalize well to real test administrations.

One possible way to overcome these potential problems in educational and psychological measurement is to conduct either a post-hoc (PH) or hybrid (H) simulations. In post-hoc simulations, instead of generated item responses, real item-response vectors, gathered from either a paper-pencil test or operational adaptive test, are used to run simulations (Thompson & Weiss, 2011). However, due to time and cost, it is not easy to obtain real responses to a large number of test items. The hybrid simulations use a combination of real and simulated item responses when running the simulation thus it can be a remedy for this problem. In hybrid simulations, based on the available responses, the missing ones are filled, then analysis is done using the hybrid data. According to Thompson and Weiss (2011), hybrid or post-hoc simulations may produce the closer results to the real adaptive testing applications than fully-generated data. Hence, more attention should be given to these two types of simulations.

Post -hoc simulations (see Bulut & Kan, 2012; Kalender & Berberoglu, 2017; Weiss & Gibbons, 2007) and hybrid simulations (Nydick & Weiss, 2009) have been highly conducted in the area of CAT. Even though, the MST is a new trend in today's testing world, to the best of my knowledge, the MST literature, does not contain such a study that illustrates an application of post-hoc and hybrid simulations. This study aims to conduct the two simulation studies as post-hoc and hybrid simulations, and compare the MST configurations under varying test length conditions. The goal of this unique study is to explore the precision of the test outcomes across the same group of students under different simulation methods, and compare the results of different simulation techniques with another, as well as to discuss the strengths and weaknesses of each method.

2. Theoretical Framework

In post-hoc simulations, real response vectors (usually collected via a paper-pencil administration) are used to run simulations. As detailed in Nydick and Weiss (2009), first the item parameters are estimated by using a

real response matrix. Then, person parameters (e.g., ability estimates) are estimated based on the estimated item parameters. Finally, in order to test specified study conditions, a PH simulation is conducted using the real response vectors. The person parameters, calculated in the full-scale (e.g., original whole data) estimates, are treated as true person parameters, and then simulation-produced person parameters are compared with those true person parameters. The PH simulations are similar to MC simulations but differ in that an actual response matrix is used to run simulations. In order to do this, a response vector including responses to all items must be available for all examinees (Thompson & Weiss, 2011). This means that all examinees included in the PH analysis must take all items in the bank. The broad research question in PH simulations is that what happens if these examinees took the test on these items (Nydick & Weiss, 2009).

PH simulations can be reflective for the operational studies; however, administering all items to all test takers is fairly difficult. This is because there is always the likelihood that some of the students do not attend some of the test administrations, and this results in not having reached items (e.g., not administered) in the data. Instead of removing, hybrid simulations can be invaluable to keep those people in the analysis. In hybrid simulations, an item bank is divided into several test forms that include common items and, each of the test form is administered to different groups of students. Then, item response matrix including actual responses is analyzed for estimating item and person parameters. After then, based on the estimated person parameters (e.g., theta), responses to not administered items are imputed, and a new response matrix having actual and simulated responses is obtained. Finally, the response data is re-analyzed for estimating item and person parameters. The broad research question in H simulations is that what happens if all examinees took all test items (Nydick & Weiss, 2009).

3. Research Questions

1. How will the test outcomes be impacted when the test length and MST design are varied in the PH simulation study?
2. How will the test outcomes be impacted when the test length and MST design are varied in the H simulation study?
3. How will the interpretations, derived from the test outcomes, be impacted across the PH and H studies under the same test length and MST design conditions?

4. Method

4.1. Building Item Pool

In this study, I first created eight test forms that included 8th grade math items. The test forms were designed so as to be essentially parallel in terms of difficulty and content balancing (e.g., numbers, algebra, equations etc.). There were 25 multiple choice (aka dichotomously scored) items in each form, and for a total of 200 items. Then, I administrated each test form to the same group of 652 students from 11 elementary schools in the 2018-2019 academic school year. After obtaining the response vectors, I first checked the dimensionality of the data separately to ensure I meet with the unidimensionality assumption for each form. Specifically, I ran the confirmatory factor analysis (CFA) with no correlated residuals for any pair of items. The CFA provided either perfect or acceptable fit for each subset. After that, I fit the whole data with the Rasch (Rasch, 1960) model to estimate item (e.g., difficulty) and person (e.g., ability scores) parameters by using concurrent calibration method (Nydick & Weiss, 2009). Finally, I removed 12 items that did not fit with the model (i.e., due to extremely high or low difficulty parameter), retained 188 items in the item bank, and re-run analyses with those remaining items.

The difficulty parameters of the retained items ranged from -1.88 to 1.42, with the mean of -0.23 and standard deviation of 0.61. The theta estimates of full scale were ranged from -2.13 to 3.14 with the mean of -0.33 and standard deviation of 1.08. Test information function for the original item bank with 188 items was given on Figure 2-a.

4.2. Multistage Testing Design

The manipulated study conditions were MST design configuration with six levels and total test length with three levels. The six levels of configuration include 1-2, 1-3, 1-2-2, 1-3-3, 1-4-4 and 1-5-5 designs. The three levels of test length include 24-item, 36-item and 48-item. All manipulated conditions were fully crossed with one another. This resulted in 18 conditions in each simulation method. Regardless of the MST design and test length, the modules always had equal numbers of items. For example, in 36-item and 1-4-4 design condition, each module is comprised of 12 items, and the panel is comprised of 108 items (12 items x 9 modules).

The model used in this study is Rasch model (Rasch, 1960) which is commonly used in operational adaptive testings (Barnard, 2018) and defines the probability of getting item *i* correct for person *p* (*Xip*=1) as;

$$P(Xip = 1|\theta p) = \frac{e^{1.7(\theta_p - b_i)}}{1+e^{1.7(\theta_p - b_i)}}, \quad (1)$$

where b_i is the item difficulty for item *i*, and θ_p is the latent ability score for person *p*.

Routing method is one of the most important elements of MST, and used to select the next module during the MST administration. In this study, I used maximum Fisher information (MFI) (Lord 1980; Thissen & Mislevy 2000) as the routing method. The item and test information function is essential for this method. MFI method calculates item level and module level (e.g., test level information) from the Equation 2, and selects the next module that maximizes the information for the current ability estimate. The item information for item *i* is calculated as;

$$I_i(\theta) = P_i(\theta)Q_i(\theta), \quad (2)$$

where θ is the ability level, $P_i(\theta)$ and $Q_i(\theta)$ are the probability of getting item *i* correct and incorrect, respectively. The total test information for a module (I_T) is the sum of the item information in a module and defined as;

$$I_T(\theta) = \sum_{i=1}^{N} I_i(\theta)_i, \quad (3)$$

where *N* is the total number of items in a module. This equation is used to run automated test assembly when building modules in all MST designs.

The automated test assembly was completed in IBM CPLEX (ILOG, Inc., 2006). First, items at similar difficulty levels were clustered into different modules, and then modules were assigned to the panel. Due to the limited number of items, one panel was built. In the two simulation studies, the items in routing modules were chosen from medium difficulty items (e.g., items that maximize information function at theta point of 0). The peaked information function points for the modules in other stages were varied from -1 to 1. Depending on the design, the difficulties of other modules were evenly distributed in this range. One can refer to Table 1 for the difficulty levels of the modules in all stages across the different MST designs. For the illustration purposes, the plots of module level information functions in 1-5-5 MST design under 48-item test condition across the post-hoc and hybrid simulations were provided in Figure 3. The plots across all specific study conditions are available upon request. The two simulations were completed in R version 2.1.1 (R Development Core Team, 2009-2015).

4.3. Study 1: Post-Hoc Simulation

This study was conducted with 263 examinees who attended all paper-pencil administrations. This means that any student that took less than eight test forms was removed from the post-hoc simulations. Item parameters were the estimates obtained in the full scale analysis (e.g., whole data estimates). The ATA was run with those item parameters to build the panel. Similarly, the true abilities of 263 people were the estimates in the full scale analysis. The real item responses belonging to 263 students were used to run varied conditions. The ability estimates calculated in the post-hoc simulations were compared with full scale ability estimates.

4.4. Study 2: Hybrid Simulation

This study was conducted with the same 263 students who attended all paper-pencil test administrations. After full scale estimates across the 652 students, all missing responses were imputed by using 'mice' package (Buuren & Groothuis-Oudshoorn, 2011) in R program. Then, a new response vector (e.g., imputed data) consisting of both real and imputed responses were re-analyzed to estimate item and person parameters. The difficulty parameters of the items ranged from -1.80 to 1.17, with the mean of -0.29 and standard deviation of 0.55. The theta estimates of hybrid data were ranged from -1.79 to 3.08 with the mean of -0.36 and standard deviation of 0.98. Test information function for this item bank with 188 items was given on Figure 2-b. The ATA was run with those new item parameters to build the modules in the panel. The person parameter estimates of the imputed data were treated as true abilities, and the ability estimates in the hybrid simulations were compared with those true abilities. For a fair comparison, the results of the same group of examinees in the hybrid simulations were compared with the those in the post-hoc simulations.

4.5. Evaluation Criteria

I evaluated the results of the study using the criteria as suggested and used in similar studies (see Bulut & Kan, 2012; Weiss & Gibbons, 2007). I calculated the following statistics: 1) absolute bias, 2) root mean squared estimate (RMSE) and 3) Pearson correlation coefficients between the true abilities and simulation-produced estimates. These statistics were computed from the following formulas.

$$\text{Absolute Bias} = \frac{\left|\hat{\theta}_j - \theta_j\right|}{N}$$

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

$$\text{Pearson Corelation} = \frac{Cov(\hat{\theta}_j, \theta_j)}{\sigma_{\hat{\theta}_j} * \sigma_{\theta_j}}$$

where N is the number of examinees, $\hat{\theta}_j$ and θ_j are the estimated and true ability parameters, and $\sigma_{\theta j}$ and $\sigma_{\theta j}$ are the standard deviations of the estimated and true ability parameters, respectively.

5. Results

5.1. Results of Post-Hoc Simulation Study

The results of absolute bias, RMSE and Pearson correlations across all study conditions obtained in post-hoc simulations were given in Table 2. The graphical illustrations of these outcomes were provided in Figure 4 (see left side). The first finding was that as the test length increased, absolute bias and RMSE decreased and pearson correlations increased with a few exceptions (e.g., 1-2-2 design & 24-item vs. 36-item and 1-5-5 design & 24-item vs. 36-item). The effect test length was more obvious under 1-2-2 and 1-3-3 MST designs and less obvious in 1-5-5 MST design. The second finding was that under the same test length condition, the lowest absolute bias and RMSE and highest correlation were always found in 1-5-5 MST design. Another finding was that the outcomes and, hereby, interpretations differed across the different MST designs. For example, under 24-item test length condition, both 1-2 and 1-3-3 MST designs produced the same amount of absolute bias (e.g., 0.63) and RMSE (e.g., 0.78). However, when the test length increased, the differences in both outcomes between the two MST designs became more obvious. This means that as the test length increased the advantage of MST design complexity was clearly seen. Overall, both test length and the type of MST design played important roles on the outcomes; however, the effect of MST design was more obvious under higher test lengths.

5.2. Results of Hybrid Simulation Study

The findings were similar to those obtained in post-hoc simulations. The first finding was that as the test length increased, absolute bias and RMSE decreased with a few exceptions (e.g., 1-2-2 design & 24-item vs. 36-item). The second finding was that under the same test length conditions, the 1-5-5 MST design always outperformed the other MST designs. The third finding was that the 1-2 MST design was the least affected and 1-3-3 design

was the most effected designs from the changes in the test length. In terms of pearson correlations, there were similar interpretations, and the benefits of increasing test length were more clearly seen.

6. Discussion and Limitation

Today, simulation techniques have become an integral part of psychometrics. There are different types of simulation methods used to test desired conditions; however, when the term of simulation is used, monte-carlo simulation comes to mind. This is because monte-carlo simulations are the dominantly used method by researchers and practitioners in relation to educational and psychological measurements. Whereas, simulation studies are not limited to the MC simulations. The purpose of this study was to provide a methodological discussion on the two types of simulation methods, and to show how to run post-hoc and hybrid simulations in the multistage testing environment, and to discuss the findings and interpretations derived from each simulation method.

The results of the study showed that in terms of absolute bias, RMSE, and pearson corrrelations, H and PH simulations resulted in comparable outcomes. The reason was that a small proportion of the data used to run H simulations was comprised of simulated data (i.e., 18%). However, in typical H simulation studies, the hybrid data consists of a greater proportion of generated data (e.g., up to 87%) (see Nydick & Weiss, 2009). There is a tradeoff between having more or less amount of generated data in H simulations. Having a lower proportion of generated data would make the data more realistic, and may yield less biased outcomes due to having more responses to higher number of items (Nydick & Weiss, 2009). The opposite would add additional source of error to the original data, however, having more proportion of real data is not always practical due to time, cost and test fatigue.

In both types of simulation studies, as the complexity of MST design increased, better results were obtained, in general, and the 1-5-5 MST design was the most recommended design. This means that having more stages, and more number of modules in the stages improved the measurement accuracy.

In the two simulation studies, as the test length increased, the outcomes were better, in general. However, some exceptions occurred in both simulation methods. This was likely due to having limited number of items in the item bank. Having more (and psychometrically good) items would have improved the automated test assembly decisions, and the effect of increasing test length would have been more obvious.

This study does not argue that MC simulations are incorrect or should no longer be used in multistage testing studies. Instead, the study argues that all three types of simulations have unique benefits and practical advantages in the implementation of operational studies. For example, in PH or H simulations studies, the quality of real data is always a concern. This is because test fatigue, the motivation of the test taker, test cheating, improper estimations for item and person parameters is always a potential risk in such studies. Since these potential risks would add additional source of random error to the test outcomes, and lead to uncertainty in the predictions (Nydick & Weiss, 2009). Hence, the interpretations derived from real or partially real data studies may not be generalized over the operational MST studies in educational and psychological measurement. Furthermore, MC simulations allow simultaneous testing of a variety of different scenarios with no cost. Considering the cost to be spent for writing test items in today's market, this advantage is enough to make MC the heart of simulations in the field of psychometrics.

PH and H studies also have their own advantages. MC simulations assume ideal conditions (e.g., a person with the ability level of θ is going to get item i correct). However, in real applications, practitioners do not always have the expected response vectors, and missing responses are inevitable. These would make it difficult to draw valid interpretations. Combining the advantages and disadvantages of the two simulation methods, even though the precision of outcomes were very comparable with PH study, the hybrid simulations can be recommended to run an MST simulation for making decisions in an operational MST. This is because in H simulations, a greater number of real examinees was used than PH studies which would make the study more powerful. As illustrated in Nydick and Weiss (2009) that H simulations with up to 80% imputed data can be as efficient as the outcomes obtained with the other simulations. Furthermore, H simulations can be a good solution for the practical limitations of PH studies such as examinee drop out, fatigue and time.

The ultimate goal of the study was to provide a unique comprehensive discussion on how different simulation techniques that could be decided and implemented in the MST environment, and the strengths and

weaknesses of each method. The aim was to explore whether interpretations and recommendations were always the same within each simulation study. Even though the proportions of complete responses were different in both studies, the interpretations were quite similar. However, post-hoc simulations produced those results with a less amount of effort and cost.

In this study I used maximum Fisher information method as the routing method. A future study can be conducted by using number-correct method or D-scoring (Han, Dimitrov, & Al-Mashary, in press) as the routing strategy. It is important to note that in traditional hybrid simulations, each examinee takes a single test form, and responses to all remaining items are imputed (Barnard, 2018). This means that a huge proportion of the data is imputed, and a small proportion is actual data. However, many of the students attended more than five paper-pencil test administrations. Hence, 18% of the responses in the data were comprised of imputed data, and 72% was comprised of real item responses. As detailed in the previous sections, compared to traditional H simulations, relatively less proportion of imputed data was used in H simulations, and this lead to obtaining similar findings with PH simulations. A further H study should be replicated with less proportion of real data and a higher proportion of imputed data. Lastly, the item bank in this study contained 188 items. Another study should be conducted with different item bank sizes.

References

Barnard, J. J. (2018). From simulation to implementation: Two CAT case studies. *Practical Assessment, Research & Evaluation*, *23*(14), 2. Retrieved from: http://pareonline.net/getvn.asp?v=23&n=14

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In R.L. Brennan (2011) Generalizability theory and classical test theory. *Applied Measurement in Education*. *24*(1), 1-21.

Bulut, O., & Kan, A. (2012) Application of computerized adaptive testing to entrance examination for graduate studies in Turkey. *Egitim Arastirmalari-Eurasian Journal of Educational Research*, *49*, 61-80.

Buuren, S. V., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, 45(3). doi: 10.18637/jss.v045.i03. retrieved from https://www.jstatsoft.org/article/view/v045i03

Davey, T., & Y.H. Lee. (2011). Potential impact of context effects on the scoring and equating of the multistage GRE Revised General Test. (GRE Board Research Report 08-01). Princeton, NJ: Educational Testing Service.

de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory.* NY: The Guilford Press.

Feinberg, R. A., & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice*, *35*(2), 36-49.

Han, K. T., Dimitrov, D. M., & Al-Mashary, F. (in press). Developing Multistage Tests Using D-Scoring Method. *Educational and Psychological Measurement*, doi: 0013164419841428.

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice*, *26*(2), 44-52.

ILOG. (2006). ILOG CPLEX 10.0 [User's manual]. Paris, France: ILOG SA.

Kalender, I., & Berberoglu, G. (2017). Can computerized adaptive testing work in students' admission to higher education programs in Turkey? *Kuram ve Uygulamada Eğitim Bilimleri*, *17*(2), 573-596. doi: 10.12738/estp.2017.2.0280. Retrieved from http://www.estp.com.tr/wp-content/uploads/2017/02/2017.2.0280.pdf

Liu, O. L., Bridgeman, B., Gu, L., Xu, J., & Kong, N. (2015). Investigation of response changes in the GRE revised general test. *Educational and psychological measurement*, *75*(6), 1002-1020.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Luecht, R. M. & Sireci, S. G. (2011). A review of models for computer-based testing. Research Report RR-2011-12. New York: The College Board.

Luecht, R. M. (2000, April). *Implementing the computer-adaptive sequential testing (CAST) framework to mass produce high quality computer-adaptive and mastery tests*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Luecht, R. M., & Nungester, R. J. (1998). Some Practical Examples of Computer-Adaptive Sequential Testing. *Journal of Educational Measurement*, *35*(3), 229-249.

Nydick, S. W., & Weiss, D. J. (2009). A hybrid simulation procedure for the development of CATs. In D. J. Weiss (Ed.), Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing. Retrieved from www.psych.umn.edu/psylabs/CATCentral/

Patsula, L. N. (1999). *A comparison of computerized adaptive testing and multistage testing* (Order No. 9950199). Available from ProQuest Dissertations & Theses Global. (304514969)

R Development Core Team. (2013). *R: A language and environment for statistical computing, reference index* (Version 2.2.1). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute of Educational Research.

Sari, H. I., & Raborn, A. (2018). What Information Works Best?: A Comparison of Routing Methods. *Applied Psychological Measurement*, *42*(6), 499-515.

Sari, H.I., & Huggins-Manley, A.C. (2017). Examining content control in adaptive tests: Computerized adaptive testing vs. computerized adaptive multistage testing. *Educational Sciences: Theory and Practice, 17,* 1759-1781

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 101-133). Hillsdale, NJ: Lawrence Erlbaum.

Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation*, *16*(1). Retrieved from https://pareonline.net/getvn.asp?v=16&n=1

Weiss, D. J., & Gibbons, R. D. (2007). Computerized adaptive testing with the bifactor model. In D. J. Weiss (Ed.). Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing. Retrieved from http://iacat.org/sites/default/files/biblio/cat07weiss%26gibbons.pdf

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

Zenisky, A. L. (2004). *Evaluating the effects of several multi-stage testing design variables on selected psychometric outcomes for certification and licensure assessment* (Order No. 3136800).

Table 1. Peaked Information Function Points Across The MST Configurations

Design	Routing	Stage 2 Modules (Easiest to Hardest)	Stage 3 Modules (Easiest to Hardest)
1-2	θ_1=0	θ_1=-1.0, θ_2= 1.0	-
1-3	θ_1=0	θ_1=-1.0, θ_2=0, θ_3= 1.0	-
1-2-2	θ_1=0	θ_1=-1.0, θ_2=1.0	θ_1=-1.0, θ_2=1.0
1-3-3	θ_1=0	θ_1=-1.0, θ_2=0, θ_3= 1.0	θ_1=-1.0, θ_2=0, θ_3= 1.0
1-4-4	θ_1=0	θ_1=-1.0, θ_2=-0.50, θ_3=0.50, θ_4= 1.0	θ_1=-1.0, θ_2=-0.50, θ_3=0.50, θ_4= 1.0
1-5-5	θ_1=0	θ_1=-1.0, θ_2=-0.5, θ_3=0, θ_4=0.5, θ_5=1.0	θ_1=-1.0, θ_2=-0.5, θ_3=0, θ_4=0.5, θ_5=1.0

Table 2. Results of Post-Hoc Simulation

Design	Absolute Bias			RMSE			Pearson Correlation		
	24 item	36 item	48 item	24 item	36 item	48 item	24 item	36 item	48 item
1-2	0.63	0.56	0.53	0.78	0.70	0.68	0.88	0.90	0.91
1-3	0.58	0.54	0.44	0.68	0.67	0.54	0.89	0.90	0.93
1-2-2	0.57	0.66	0.50	0.70	0.81	0.60	0.87	0.88	0.92
1-3-3	0.63	0.52	0.41	0.78	0.62	0.52	0.88	0.92	0.93
1-4-4	0.50	0.52	0.40	0.60	0.62	0.51	0.90	0.92	0.93
1-5-5	0.45	0.44	0.38	0.57	0.61	0.47	0.91	0.93	0.94

Table 3. Results of Hybrid Simulation

Design	Absolute Bias			RMSE			Pearson Correlation		
	24 item	36 item	48 item	24 item	36 item	48 item	24 item	36 item	48 item
1-2	0.63	0.54	0.53	0.78	0.68	0.67	0.88	0.90	0.91
1-3	0.56	0.57	0.45	0.68	0.69	0.55	0.89	0.89	0.93
1-2-2	0.54	0.63	0.47	0.66	0.78	0.58	0.88	0.88	0.93
1-3-3	0.64	0.53	0.44	0.79	0.63	0.55	0.88	0.92	0.93

1-4-4	0.51	0.46	0.46	0.62	0.58	0.56	0.89	0.92	0.93
1-5-5	0.45	0.48	0.35	0.56	0.60	0.44	0.91	0.92	0.94

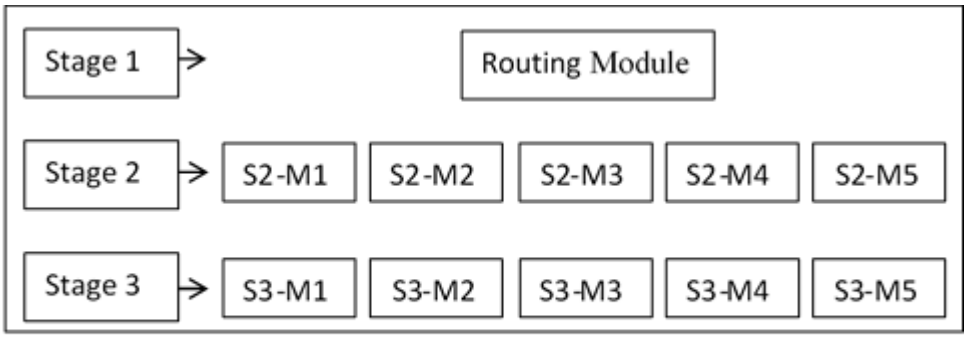


Figure 1. An example of 1-5-5 MST panel design
Note: S=Stage, M=Module

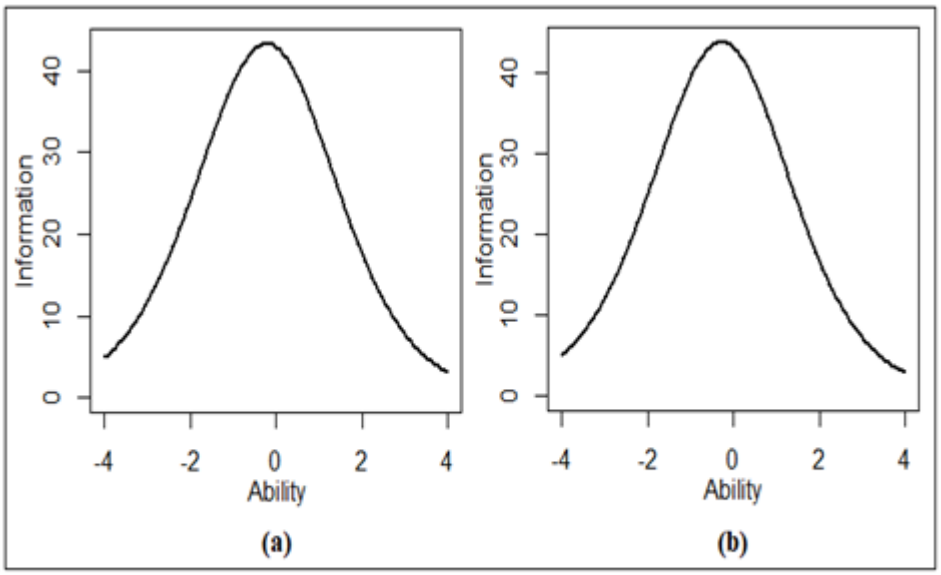
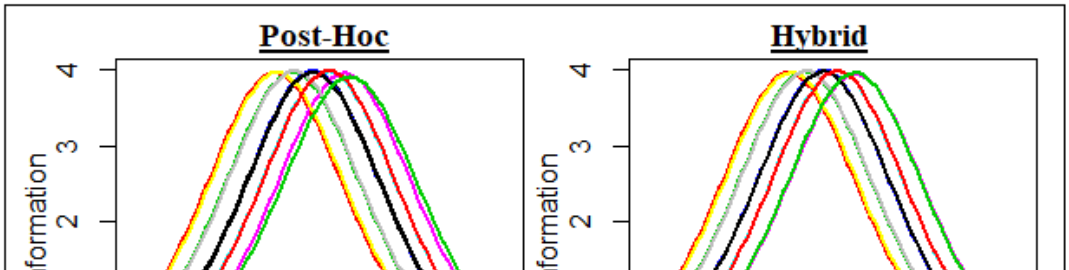


Figure 2. Test information function plots for the original (a) and hybrid simulation (b) item banks.



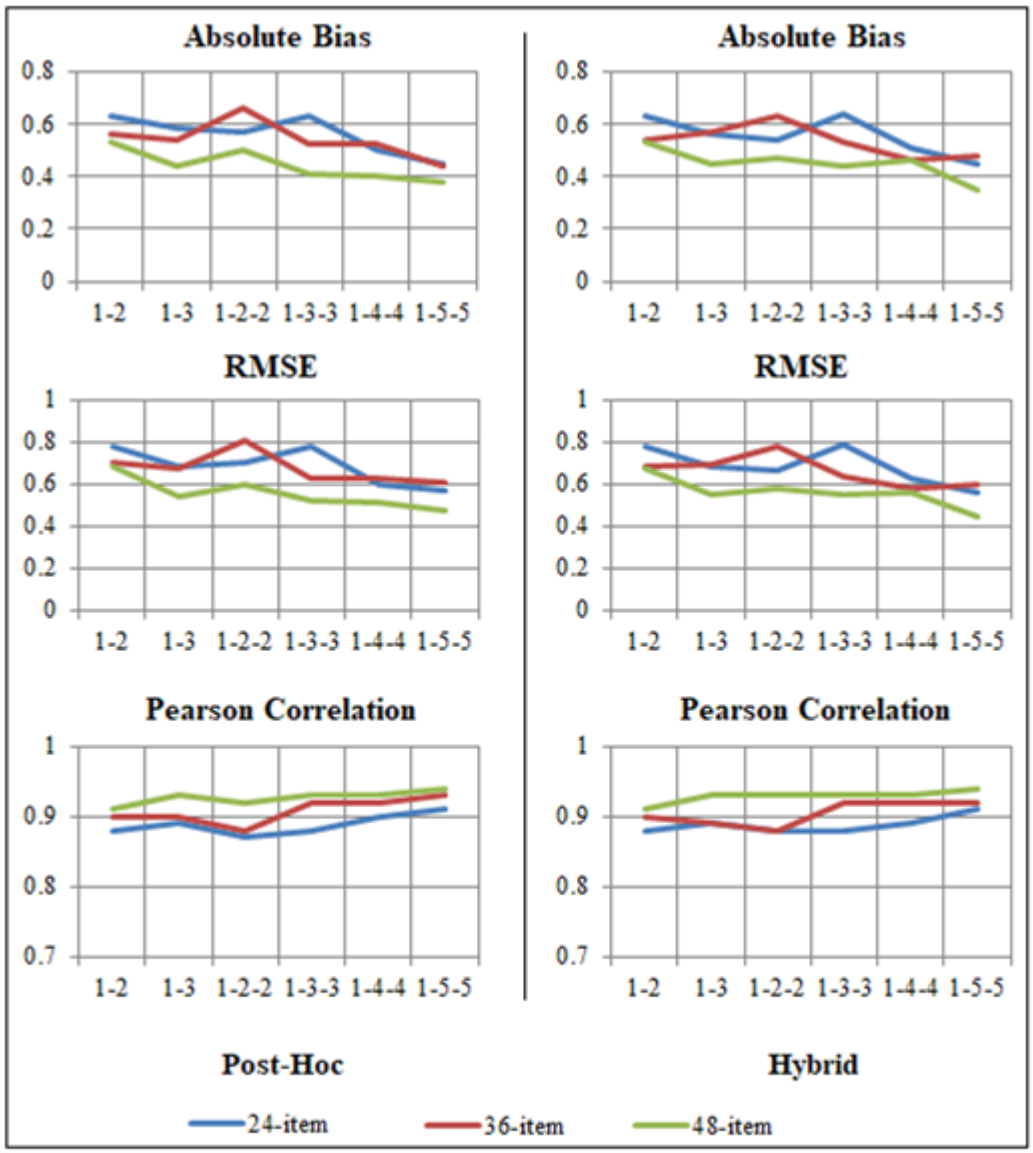


Figure 4. Graphical illustrations of absolute bias, root mean square error (RMSE) and person correlations across the different MST designs and test length conditions in post-hoc and hybrid simulations.