Validation of Setting and Design of Multi-Stage Testing (MST) to Portray Students' Achievement on Reading Literacy in AKMI 2021

Wahyu Widhiarso[1,*] Ali Ridho[2]

[1] *Universitas Gadjah Mada, Yogyakarta, Indonesia*
[2] *Universitas Islam Negeri Maulana Malik Ibrahim Malang, Malang, Indonesia*
[*]*Corresponding author.* Email wahyu_psy@ugm.ac.id

ABSTRACT
This study identifies the appropriateness design in measuring student abilities through multi-stage testing (MST) in the AKMI program. This study is a follow up of the previous study which carried out the calibration process of the items after the try-out process. Items that meet the qualifications are then included in the item bank which is the material for the implementation of the MST. This research conducts a simulation study to identify the effectiveness of the MST design that has been developed as well as alternative designs. The input used in the simulation process is the distribution of students' abilities and items on the question bank and the MST design. The first two aspects were obtained from the preliminary study while the MST design was obtained from the test administrator of AKMI. The results of the analysis show that the MST design with three stages and the scoring procedure using the classical model can produce student score information with high precision.

Keywords: *Multi-stage testing, AKMI, Students' ability.*

1. INTRODUCTION

The use of multi-stage testing (MST) in many practical settings was popular [1] – [3] but studies that maintain this topic still rare although many experts developed various methods for administering tests using MST. There are two reasons to answer question why only small research already examined implementation of MST in practical setting. First, implementation of MST requires complex structure [4] and infrastructure (e.g., item bank, systems) especially when it operate in large scale assessment [5]. As consequently, little data exist to support study on MST. Second, conducting MST is recent approach that altered traditional strategies so that it might affected by psychological status of examinees. Hence, successfulness of MST in large setting assessment also depends on readiness of examinees to participate in new assessment procedure.

MST is already developed many years ago [6] as one types of procedures proxy to computerized adaptive tests (CAT). The use of item response theory in practical setting combined with supported technology (e.g., computers) fostered the potential to develop more effective. Means that developed test were shorter but tends to obtain greater score reliability. Nowadays, there are different varieties types of MST exist [7] and employed in many practical settings. We use one approach of these procedures in Assessment Competencies of Madrasah (AKMI). An instrument to be administered for practical purposes requires a validation process in order to produce accurate information about the individual being tested. MST is a new computerized test delivery technology aimed at enhancing the quality of credentialing exams. MST offers the potential to increase testing efficiency and accuracy compared to typical test with fixed test length. The purposes of this paper is to report result of validation of MST on AKMI because our goal writing this paper is giving some insights for MST developer to validate their instrument and procedures.

There are several issues that need to be resolved regarding our MST design in AKMI. There are various factors affect the quality of measurement of the test in each stage (e.g., testlets) set of items in each panel and the scores that result from this type of MST design.

First, sample sizes in MST implementation. [8] examines the sample sizes needed for MST. They explore

effects of sample sizes of different item calibration on person ability and classification accuracy. Using a Monte Carlo simulation, they found that samples of approximately 300 and 1,000 resulted in similar theta estimates and decision accuracies. AKMI was followed by thousand examinees so it possible that problem related to this issue was small. Second, the size and characteristics of the item bank. Characteristics of item parameter existing in item bank, for example distribution of item difficulty and discrimination level of items affect the breadth of test information that can be achieved across ability level of examinees [9]. When the breadth of the test is to narrow there will be area on the continuum of the ability level that will produce lower test information. This is because number of items that effectively discriminating examinee in certain ability were small due to limitations based on the size and characteristics of the item bank. The number of items per stage (e.g., testlet size) can directly affect the capability of the panel to adapt to examinees having more extreme abilities. Hence, number of items per stage should be developed according to the number of items needed per panel.

Third, scoring system. In MST scoring, test administrator might select one computational scoring among various techniques [10]. Three are several types of scoring MST available, score on individual items level, cumulative score of the panel and final score. Fourth, MST typically administered using set of items (e.g., item bundles, testlets) rather than individual item. Although this is one advantage of MST because the use of set of items will encounter some problems regarding multidimensionality of test score, local independence and negative variance due to overlapping content to measure, in practice there are still specific issues that need to be resolved.

This study validate the MST settings implemented in student assessment in the AKMI 2021 program. Researchers and administrators of the 2021 AKMI have developed certain designs that need to be validated in terms of quality and accuracy for measuring students' competencies. Information about the quality and accuracy is important to identify so that all the concerns of AKMI, schools and students can be fulfilled properly.

2. METHODS

2.1 Research Design

The purpose of this study was to validate and identify the effectiveness of the MST design on the AKMI 2021. The analysis procedure was conducted using several steps according following structure (see Figure 1 for illustration). (1) Identify the distribution of students' abilities in the population based on preliminary studies (field testing and first phase validation). Before this study, we already conduct a preliminary study to identify and map students' abilities. This study found information regarding the distribution (mean & standard deviation) of students' abilities by Item response theory (IRT) modeling and scoring. This activity provides information of students' abilities, called theta using expected a posteriori (EAP) scaled score for each competency to measure. After the mean and standard deviation are obtained, the next step is generating data (e.g., 100,000 students) according to that information to estimate population distribution regarding competency to measured. One of the information that need to be inputted on the simulation system is data regarding students score so that we try to fulfilled this requirement. On the other word, the purpose of this identification is to support the process of simulating the implementation of MST because one of the information needed to input is the distribution of the abilities of students who participate in MST.

(2) Identify psychometric properties in the existing item bank for each measured competency. The field-testing phase carried out in the previous study has selected the hundred items that have been field tested and calibrated before compilated in the item bank. There are several criteria used to enter items into the item-bank. For example, items with high discriminatory power and have good model fit with the IRT model used in the scoring procedure through the developed MST. Thus, all the items in item bank have equal qualities so that an automated test assembly (ATA) process is possible to implement. (3) Developing MST design that contains the rules and procedures for the implementation up to the scoring. In this study, the developed MST has three stages, starting with the initial stage of identifying students' general abilities followed by two consecutive stages. The score obtained at this early stage determines which path and stage that students will take. (4) Implementing a simulation of the application of MST based on the information entered (i.e., item bank, ability distribution, MST design) which results in a score for each person participating in the MST. (5) Estimating the correlation between student scores that have been determined from the start through the data generation process and student scores resulting from their participation when following the MST with a predetermined design. The higher the correlation obtained, the more precise the MST design developed. The high correlation also indicates that additional findings regarding adequacy of items on the item bank or the appropriateness of the number of students participates in the MST.

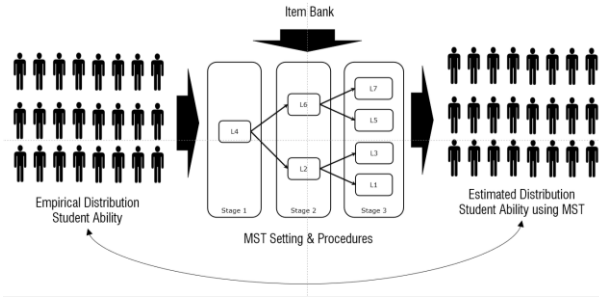

Figure 1 Design of this research.

2.2 Participant

The participants of this study were students from schools under the coordination by Ministry of Religious Affairs (MORA) who participated in the AKMI program. Participants come from various school levels, namely elementary level (Ibtidaiyah), Junior High School (Tsanawiyah), Senior High School (Aliyah). This paper will only examine the implementation of MST in the AKMI program followed by students at the Ibtidaiyah level.

2.3 Instruments

The research instrument consists of four tests that measure different constructs, namely reading, numeracy, science and socio-cultural competencies. MST was held for each competency so that all competencies measured have specific item bank. All test that assess four competencies have been validated using qualitatively approach (e.g., expert judgment) and quantitatively approach (e.g., structural validity), both at the item level (e.g., item discrimination) and test level (e.g., reliability). Psychometric properties of calibrated items was obtained from implementing item try-out and field-test from total 3,690 students.

2.4 Analysis Procedure

Data analysis in this study was conducted out by examining the results of each MST implementation using simulation study that already carried out [11]. In each simulation process, there are three inputs of information: distribution of students' abilities, list of items in the item bank and the MST design. The two information in the beginning were fixed because that information was attained from previous studies whereas the MST design is varied. This study examines effectiveness of various MST designs. There are several considerations that can be used to develop the MST design. This consideration is a trade-off between two aspects: practicality and precision. If MST design are merely focused on the practical aspect, the precision aspect will be lost, and vice versa. Therefore, this study analyzed data descriptively by comparing the effectiveness of several MST designs that were tested in the form of simulations

3. RESULTS AND DISCUSSIONS

3.1 Results

3.1.1 Distribution Item Parameter of Item Bank

Figure 1 describes the distribution of item difficulty compared to distribution of students' abilities of four competencies tests (reading, numeracy, science and social culture). That figure presents two distributions which is called as item difficulty distribution (on the top) and student abilities distribution (at the bottom). This information is come from different entity, but it can be combined on the one continuum as they have been scaled on the same metric, namely the logit scale. Analysis shows that the item bank for each test closed to an ideal distribution because item difficulty of item bank covers wide abilities from low to high levels. Another interesting point of this finding is the range of the distribution of item difficulty is broader than the distribution of students' abilities. It means that the items in the item bank have capability to discriminate student at various level ability, not only individual that meet characteristics of population but also students with characteristics that are not represented by the population, for example students who have very low or very high abilities.

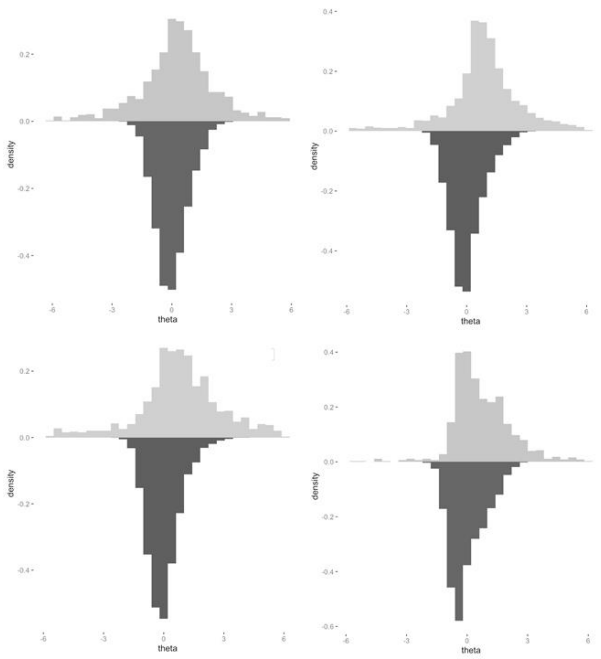

Figure 2 Item Map of Four Tests.

3.1.2 Distribution of Ability

The distribution of students' abilities in all competency are similar. All score fit with normal distribution with a mean score (theta) approximately closed to 0 as well as the standard deviation that approximately closed to 1. The breadth of the distribution

of student abilities smaller than the breadth of the distribution of item difficulty (see Figure 1).

3.1.3 Simulation Results

The simulation results of the MST designs show that the MST design with three steps has a fairly high accuracy compared to the design with more steps. The correlation between students' ability scores inputted to the MST system (empirical score) and student's scores after participate in MST (estimated score) is high. This finding shows that the MST with three stages has two advantages in terms of practicality or efficiency as well as in terms of precision.

4. CONCLUSION

This study was aimed to identify MST design that most appropriate to apply for implementation in AKMI. There are many MST alternative designs that have been tested and simulated using this study. These variations were result of combination among several aspects such as number of stages in MST, the number of items for each stage, the criteria for passing a certain path from the existing paths and finally the scoring system. This study found that the most appropriate design for conducting measurements using MST with resources (e.g., item banks) currently owned by AKMI 2021 is MST with three stages. The MST design with three stages has a very high accuracy supported by a high practicality aspect as well.

This study found that a simple MST design was sufficient to measure individual abilities with high precision. In this study, the design referred to by the simple design is the MST design with three stages. Finding of this study is supported by previous study. [12] already demonstrated that five-module design of a four-stage MST added little improvement in scoring reliability compared to the three-stage MST. These finding answers important question in MST design such as trade-offs of between adding modules and stages and maintaining the simplicity and pool use of MST. In general, their study confirms that a maximum of four modules is desirable at any one stage (for a fixed module length) and that three stages may be sufficient.

In first version, the design of MST in AKMI consists of three stages. Stage 1 is the beginning in where all student follows the test with a common set of items of moderate level of difficulty. This set is called the routing test. In each stage, student work on set of items which is called as a module or testlet. After work on stage 1, students' competencies were estimated as the basis of selecting alternative modules in stage 2. Compared to module in stage 1, module in stage 2 is more informative from a measurement point of view for each examinee. Estimated ability in stage 2 have higher precision than estimated ability in stage 1 because there is matching process between student level ability and task difficulty. In AKMI, there are several modules in stage 2 with vary on the basis of item difficulty. After stage 2 student work on stage 3 which is consists of item or task that most fit to the ability of students. Thus, result of estimation of ability in stage 3 was have highest precision compared to stage 1 and stage 2.

AUTHORS' CONTRIBUTIONS

All authors conceived and designed this study. All authors contributed to the process of revising the manuscript, and at the end all author have approved the final version of this manuscript.

ACKNOWLEDGMENTS

The researchers would like to thank AKMI (Asesmen Kompetensi Madrasah Indonesia) and Ministry of Religious Affairs Indonesia for supporting facilities this work.

REFERENCES

[1] J. Lewis, H. Lim, F. Padellaro, S. G. Sireci, A. L. Zenisky, Setting and Validating Multiple Standards on a Multistage-Adaptive Test. Educational Measurement: Issues and Practice, 2021, doi:10.1111/emip.12434

[2] D. MacGregor, S. J. Yen, X. Yu, Using Multistage Testing to Enhance Measurement of an English Language Proficiency Test. Language Assessment Quarterly, 2021, 1-22, doi:10.1080/15434303.2021.1988953

[3] H. J. Shin, K. Yamamoto, L. Khorramdel, F. Robin, Increasing Measurement Precision of PISA Through Multistage Adaptive Testing. Paper presented at the Quantitative Psychology, Cham, 2021.

[4] D. Yan, C. Lewis, A. A. V. Davier, Overview of computerized multistage tests. In D. Yan, C. Lewis, & A. A. von Davier (Eds.), Computerized multistage testing. London: CRC Press: Taylor & Francis Group, 2014.

[5] A. Zenisky, R. K. Hambleton, R. M. Luecht, Multistage testing: Issues, designs, and research. In Elements of adaptive testing, 2009, pp. 355-372.

[6] F. M. Lord, Application of Item Response Theory to Practical Testing Problems. Hillsdale, New Jersey: Lawrence Erlbaum Associates Publishers, 1980.

[7] P. Bauer, F. Bretz, V. Dragalin, F. König, G. Wassmer, Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls. Statistics in Medicine, 2016, 35(3), 325-347. doi:10.1002/sim.6472.

[8] S. C. Chuah, F. Drasgow, R. Luecht, How Big Is Big Enough? Sample Size Requirements for CAST Item Parameter Estimation. Applied Measurement in Education, 2006, 19(3), 241-255. doi:10.1207/s15324818ame1903_5

[9] L. Yang, M. D. Reckase, The Optimal Item Pool Design in Multistage Computerized Adaptive Tests With the p-Optimality Method. Educational and Psychological Measurement, 2020, 80(5), 955-974. doi:10.1177/0013164419901292

[10] K. T. Han, D. M. Dimitrov, F. Al-Mashary, Developing Multistage Tests Using D-Scoring Method. Educational and Psychological Measurement, 2019, 79(5), 988-1008. doi:10.1177/0013164419841428

[11] S. Kim, T. Moses, An Investigation of the Impact of Misrouting Under Two-Stage Multistage Testing: A Simulation Study. ETS Research Report Series, 2014, (1), 1-13.

[12] R. D. Armstrong, D. H. Jones, N. B. Koppel, P. J. Pashley, Computerized adaptive testing with multiple-form structures. Applied Psychological Measurement, 2004, 28(3), 147-164. doi:10.1177/0146621604263652