Advances in Social Science, Education and Humanities Research, volume 633
International Conference on Madrasah Reform 2021 (ICMR 2021)

Sociocultural Literacy Assessment: Validation of Multistage Generalized Partial Credit Testing Design

Ali Ridho[1,*]

[1] *Department of Psychology, Universitas Islam Negeri Maulana Malik Ibrahim Malang*
[*] *Corresponding author. Email: aliridho@uin-malang.ac.id*

ABSTRACT
This aims to (1) describe the analysis results of sociocultural literacy items based on generalized partial credit model (GPCM) and (2) validate the design of multistage testing (MST). Two field tests involving 5881 secondary madrasa students on 1416 items comprising essential competencies; nationality commitment, tolerance, anti-violence, accommodative, and inclusive. The items format developed in the literacy were multiple choices, complex multiple choices, matching, and fill-in. The results unravel that (1) 1109 items (78.3%) have good item-discrimination and (2) the MST validation result in a high accuracy estimation of the literacy.

Keywords: *Validation, Multistage testing (MST), Sociocultural literacy.*

1. INTRODUCTION

Multistage testing (MST) is a popular approach of administering tests [1-4]. Its effectivity in accommodating the balance between test content and security attract test administrators to adapt it. Although MST has superior functions, the implementation is not easy; requiring a solid collaboration amongst psychometrician, test developers, and sufficient computer networks and servers.

The Ministry of Religious Affairs of the Republic of Indonesia (MoRA) is committed to strengthen madrasah, support madrasa students' literacy, and develop the identify of madrasa through National Assessment of Madrasa Competencies (*Asesmen Kompetensi Madrasah Indonesia*, AKMI). In this case, a multistage testing (MST) is administered. There are four competencies including in the assessment; reading, science, numeracy and sociocultural. This paper addresses the sociocultural literacy of primary madrasa students' literacy.

A large items pool is one of the requirements to administer MST [5]. The spreading of the items has a range of difficulties, which accommodate all abilities; from the lowest theta to the highest one. Given the sufficient large pool, the module development in each stage will be easier.

Besides the large items pool, the assessment instrument for sociocultural literacy using MST needs validation to yield accurate information about students' sociocultural literacy. The current study aims to validate the MST design of the items. The results are expected to provide insight on the validation process for items developers who plan to do similar research.

1.1. Sociocultural Literacy

Cultural literacy as knowledge of meaning systems, and the ability to negotiate those systems within different cultural contexts [6] is adapted in the items development. Sociocultural literacy is defined as students' abilities to know, respond, reflect, evaluate, and create knowledge, behavior plans, and action plans relating to the nationality commitments, tolerance, anti-violence, accommodative, and inclusive. It is based on fields of history, sociology, anthropology, and relevant strategic issues; all of which relate to the contexts of personal, society, and religions. The aim of the literacy is to elevate students' knowledge and support their social participations.

1.2. Multistage Generalized Partial Credit Testing

The combination of MST and GPCM (Generalized Partial Credit Model) is called Multistage Generalized Partial Credit Testing (MSGPCT) in this paper. The response patterns of students' scores (0, 1, 2, 3) of their answers towards the sociocultural literacy items are subject to calibration. The measurement model used for calibration is Generalized Partial Credit Model (GPCM)

[7]. Muraki developed Master's Partial Credit Model (PCM) [8] to become GPCM as follows:

$$P_i(U_j = 1 \mid \theta) = \frac{\exp[a(\theta - b_j)]}{1 + \exp[a(\theta - b_j)]}, \quad (1)$$

untuk item-item dikotomi (0, 1); dan

$$C_{jk} = P_{jk|k-1,k}(\theta) = \frac{P_{jk}(\theta)}{P_{j,k-1}(\theta) + P_{jk}(\theta)} = \frac{\exp[a(\theta - b_j)]}{1 + \exp[a(\theta - b_j)]}, \quad (2)$$

for polytomous items (0, 1, 2, …), where the items difficulties are shown by the parameter *b* (the usual range is from -6 to 6) while the parameter *a* is discrimination-items. The items can function properly if $a > 0$.

The items with the given parameters are developed into model or booklet to be given to the test takers. The modes of administering tests have been advancing rapidly, from linear test approach (LT), computerized adaptive test (CAT), and the contemporary one is multistage test (MST) [9]. The MST design for AKMI is presented at Figure 1.

From a historical perspective, multistage testing was initiated by reference [10], then it becomes an alternative in a contemporary test administration system [11-13][4][9].

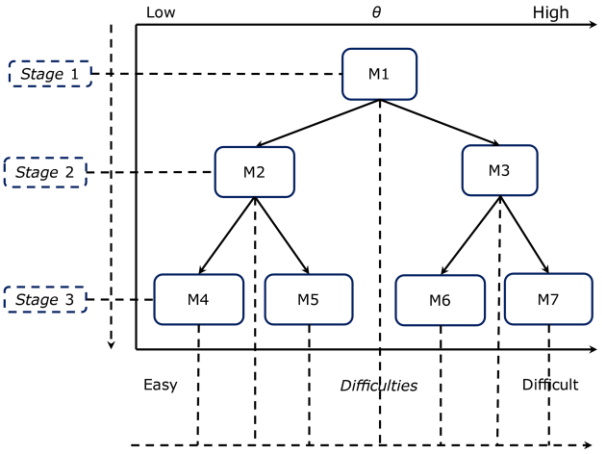

Figure 1. MST design for AKMI

MST relies on the items grouping called module. The modules will be responded by the test takers sequentially. In Figure 1, 7 modules are used for each stage. A test taker begins with module 1 (M1). If his/her ability is identified high at the stage, then at the second stage he/she will have M3. The results of M1 and M3 will determine whether he/she shall have M6 or M7. If his/her ability is high, then the next module will be M7. Otherwise, he/she will have M6. Overall, there are four possible paths encountered by the test takers; (1) M1-M2-M4, (2) M1-M2-M5, (3) M1-M3-M6, and (4) M1-M3-M7.

2. METHODS

2.1. Participants

Items discrimination parameter (*a*) and items difficulty (*b*) were obtained through a field test towards the developed sociocultural literacy test items. The field test was carried out twice involving 3325 and 2974 primary madrasah students, respectively. Thus, a total number of participants were 5881 students.

2.2. Instrument

To assess the sociocultural literacy, 1416 items were developed covering several essential competencies: (1) nationality commitment; (2) tolerance; (3) anti-violence; (4) accommodative and inclusive. The items format was: (1) multiple choices; (2) complex multiple choices; (3) matching; dan (4) fill-in.

2.3. Analysis

Considering the mixed format of the items used, generalized partial credit model (GPCM) is a representative approach [14]. Technically, the software used for data analysis was "mirt: A multidimensional item response theory package for the R environment" [15]. The results of two validations were calibrated items.

The items are called good if the items-discrimination has positive values ($a > 0$). The identified items difficulty level (*b*) was utilized to develop the MST module. Each module had a distinct difficulty level, included items that have relevant difficulty levels, and referred to the stages depicted in Figure 1.

The validation of MSGPCT design applied simulations using R package "dexterMST: dexter for Multi-Stage Tests" [16] and "mstR" [17]. The scheme of routing rule follows Figure 1 with four possible paths: (1) M1-M2-M4, (2) M1-M2-M5, (3) M1-M3-M6, dan (4) M1-M3-M7.

3. RESULTS AND DISCUSSION

3.1. Good and Bad Items

The results of items analysis applying GPCM model are presented in Table 1. Based on the evaluation of items-discrimination parameter (*a*), which equals to 0 or less than 0, the bad items were reviewed by experts to be revised by the items developers or they were not used. The revised items were then recompiled along with the non-tested items. The group of items is assembled into booklets to be tested in the second stage. Bad items in the second stage were not used to design the MST; only good items were chosen for the items pool.

Table 1. The results of items calibration using GPCM

	Total items	Good items	Bad items
Field test 1	554	306	248
Field test 2	1110	803	307
		1109	

Table 1 indicates the results of items calibration, the the field-tested items in the first and second stage, and the number of bad and good items. All accumulated items provide information on the test level in the range of theta shown in Figure 2. In details, the dispersion of the items-difficulty level is depicted in Figure 3. Both Figure 2 and Figure 3 reveal that the items dispersion has fulfilled all levels of difficulties from the lowest (easy) to the highest (difficult). Therefore, from the perspective of items pool, the number of items available meet the requirement to do MST.

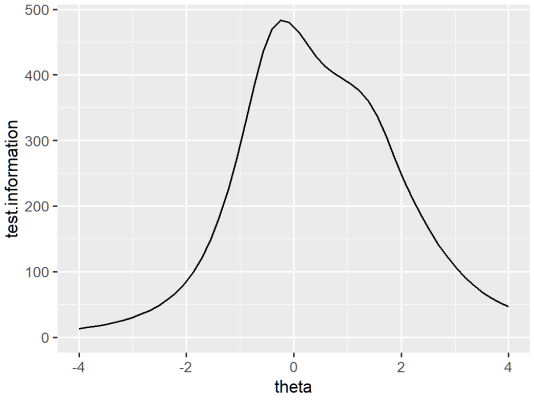

Figure 2. Test information based on 1109 items

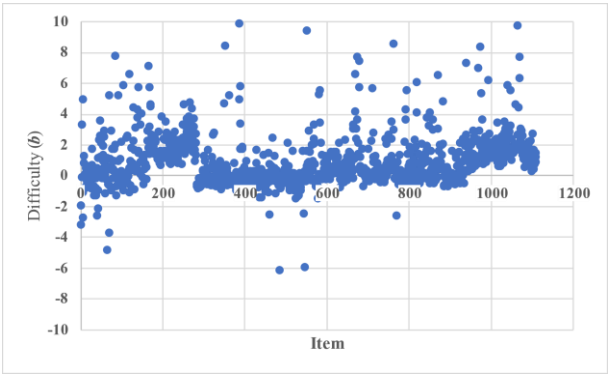

Figure 3. The dispersion of item-difficulty (*b*) for 1109 items

The natures of bad items are shown in Figure 4, while Figure 5 indicates good items. The bad items have negative or null item-discrimination; the higher the students' ability, the lower their probability to answer correctly. On the contrary, the good items have characteristics that suit the objective of sociocultural literacy assessment; the higher students' ability, the higher their probability to answer correctly.

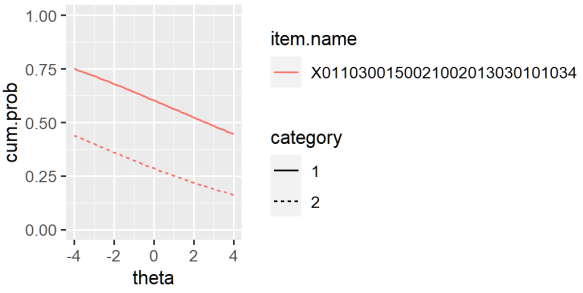

Figure 4. The sample of ICC for bad item

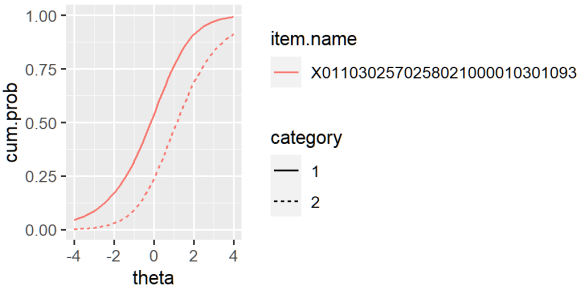

Figure 5. The sample of ICC for good item

3.2. MST Validation

The simulation to validate MST was done in a similar empirical way. The implementation of MST design (Figure 1) was simulated 500 times and the results are in Figure 6.

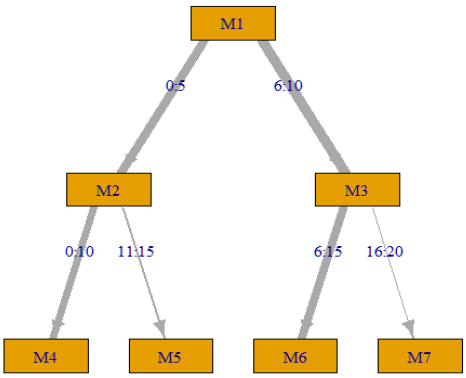

Figure 6. The path for MST simulation

The mean correlation between simulation data and theta estimation is M_r = 0.913 with SD_r = 0.003. In specific, a scatter plot is presented to compare simulated and estimated theta from the first simulation (Figure 7) and the second simulation (Figure 8). Referring to the strong correlation, the simulation reveals that the chosen MST design will be able to produce accurate estimation for madrasa students' sociocultural literacy.

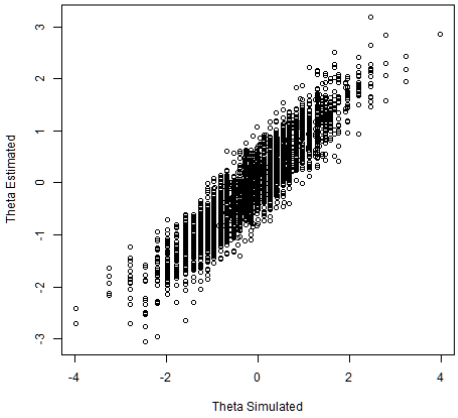

Figure 7. The 1st simulation (r = .917)

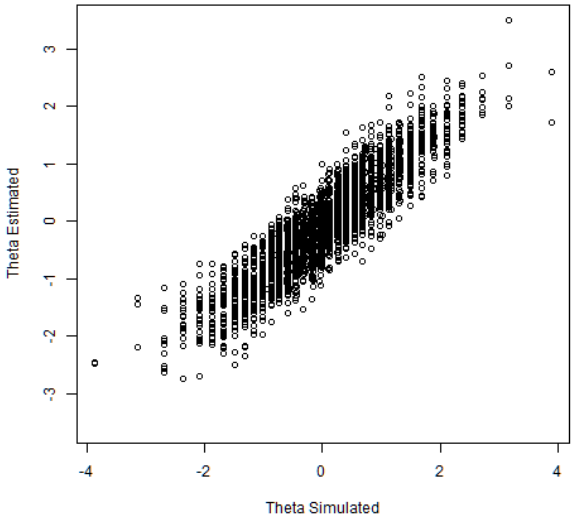

Figure 8. The 500th simulation (r = .907)

4. CONCLUSION

The number of items with good characteristics of item-discrimination are sufficiently available fulfilling the requirement to carry out MST. The chosen MST design in the simulation based on the empirical items has proved to be able to produce accurate estimation of ability parameter (theta).

ACKNOWLEDGMENTS

The author thanks the Directorate of Curricullum, Infrastructures, Institutions, and Students Affairs, Directorate General of Islamic Education, the Ministry of Religious Affairs of Republic of Indonesia.

REFERENCES

[1] L. Cai, A.D. Albano, L.A. Roussos, An investigation of item calibration methods in multistage testing, Measurement: Interdisciplinary Research and Perspectives 19(3) (2021) 163-178. DOI: https://doi.org/10.1080/15366367.2021.1878778

[2] G. Li, Y. Cai, X. Gao, D. Wang, D. Tu, Automated test assembly for multistage testing with cognitive diagnosis, Frontiers in Psychology 12(1347) (2021). DOI: https://doi.org/10.3389/fpsyg.2021.509844

[3] 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: https://doi.org/10.1080/15434303.2021.1988953

[4] H.J. Shin, K. Yamamoto, L. Khorramdel, F. Robin, Increasing measurement precision of PISA through multistage adaptive testing, in: M. Wiberg, D. Molenaar, J. Gonzales, U. Bockenholt, J.S. Kim (Eds.), Quantitative Psychology Springer Proceedings in Mathematics & Statistics, Springer, 2021, pp.325-334. DOI: https://doi.org/10.1007/978-3-030-74772-5_29

[5] L. Yang, M.D. Reckase, The optimal item pool design in multistage computerized adaptive tests with the p-optimality method, Educational and Psychological Measurement 80(5) (2020) 955-974. DOI: https://doi.org/10.1177/0013164419901292

[6] G.G. Ochoa, S. McDonald, Cultural Literacy and Empathy in Education Practice, Springer, 2020.

[7] E. Muraki, A generalized partial credit model: Application of an EM algorithm, Applied Psychological Measurement 16(2) (1992) 159-176. DOI: https://doi.org/10.1177/014662169201600206

[8] G.N. Masters, A Rasch model for partial credit scoring, Psychometrika 47(2) (1982) 149-174. DOI: https://doi.org/10.1007/BF02296272

[9] D. Yan, C. Lewis, A.A. von 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.

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

[11] A. Ariel, B.P. Veldkamp, K. Breithaupt, Optimal testlet pool assembly for multistage testing designs, Applied Psychological Measurement 30(3) (2006) 204-215. DOI: https://doi.org/10.1177/0146621605284350

[12] I.I. Bejar, E.A. Graf, Updating the duplex design for test-based accountability in the twenty-first century, Measurement: Interdisciplinary Research and

Perspectives 8(2-3) (2010) 110-129. DOI: https://doi.org/10.1080/15366367.2010.511976

[13] A.D. Mead, An introduction to multistage testing, Applied Measurement in Education 19(3) (2006) 185-187. DOI: https://doi.org/10.1207/s15324818ame1903_1

[14] R. Park, J. Kim, H. Chung, B.G. Dodd, Enhancing pool utilization in constructing the multistage test using mixed-format tests, Applied Psychological Measurement 38(4) (2014) 268-280. DOI: https://doi.org/10.1177/0146621613515545

[15] R.P. Chalmers, mirt: A multidimensional item response theory package for the R environment, Journal of Statistical Software 48(1) (2012) 1-29.

[16] T. Bechger, J. Koops, R. Zwitser, I. Partchev, G. Maris, dexterMST: dexter for Multi-Stage Tests, 2021.

[17] D. Magis, D. Yan, A. von Davier, mstR: An R package to generate multistage testing designs, 2017.

R Script

```
library(dexterMST)
library(mstR)
library(dplyr)
i =1

tmpcor <- 0.8567
db = create_mst_project("akmi")

for (i in 1:500){
 paritem <- data.frame(read.csv("param_1.csv")[,19])
 colnames(paritem) <- "measure"
 items <- data.frame(item_id=sprintf("item%02i",1:nrow(paritem)),
     item_score=1,
     delta=paritem[order(paritem$measure),])
 items<-items[sample(nrow(items), 70), ]
 row.names(items) <- 1:70
 items = data.frame(item_id=sprintf("item%02i",1:70),
        item_score=1, delta=sort(runif(70,-1,1)))
 design = data.frame(item_id=sprintf("item%02i",1:70),
     module_id=rep(c('M4','M2','M5','M1','M6','M3', 'M7'),
           each=10))
 routing_rules = routing_rules = mst_rules(
  `124` = M1[0:5] --+ M2[0:10] --+ M4,
  `125` = M1[0:5] --+ M2[11:15] --+ M5,
  `136` = M1[6:10] --+ M3[6:15] --+ M6,
  `137` = M1[6:10] --+ M3[16:20] --+ M7)
 theta = rnorm(3000, mean=-0.029, sd=0.878)
 dat = sim_mst(items, theta, design, routing_rules,'all')
 dat$test_id='sim_test'
 dat$response=dat$item_score
 scoring_rules = data.frame(
  item_id = rep(items$item_id,2),
  item_score= rep(0:1,each=nrow(items)),
  response= rep(0:1,each=nrow(items))) # dummy respons

 db = create_mst_project(":memory:")
 add_scoring_rules_mst(db, scoring_rules)

 create_mst_test(db,
     test_design = design,
     routing_rules = routing_rules,
     test_id = 'sim_test',
     routing = "all")
 add_response_data_mst(db, dat)
 design_plot(db)
 f = fit_enorm_mst(db)
 abl = ability(get_responses_mst(db), f) %>%
  inner_join(tibble(person_id=as.character(1:3000),
        theta.sim=theta), by='person_id')
 png(file=paste0("./plot/plot",i,".png"))
 plot(abl$theta, abl$theta.sim,
  xlab="Theta Simulated",
  ylab="Theta Estimated")
 dev.off()
 abl = filter(abl, is.finite(theta))
 rescor <- cor(abl$theta, abl$theta.sim)
 tmpcor <- rbind(tmpcor, rescor)
}
tmpcor
```