Eurasia Journal of Mathematics, Science & Technology Education, 2015, **11**(5), 1105-1118

Multidimensional Computerized Adaptive Testing for Indonesia Junior High School Biology

Bor-Chen Kuo
National Taichung University of Education, TAIWAN
Muslem Daud
National Taichung University of Education, TAIWAN &
Serambi Mekkah University B. Aceh, INDONESIA
Chih-Wei Yang
National Taichung University of Education, TAIWAN

•Received 18 December 2014•Revised 18 March 2015 •Accepted 20 April 2015

This paper describes a curriculum-based multidimensional computerized adaptive test that was developed for Indonesia junior high school Biology. In adherence to the Indonesian curriculum of different Biology dimensions, 300 items was constructed, and then tested to 2238 students. A multidimensional random coefficients multinomial logit model was used to develop multidimensional scales. A simulation study based on the item and ability parameters estimated from real data is conducted to evaluate the performance of the proposed Biology-MCAT system. The results show that the bias and standard error of Biology-MCAT system are acceptable. It suggests that, this online assessment system can be a model for Indonesian National Examination in future.

Keywords: computerized adaptive testing, Indonesia assessment, science education, multidimensional random coefficients multinomial logit model

INTRODUCTION

Computers have played an important role in supporting both teaching/learning (Weiss, 2011; Ozyurt & Ozyurt, 2013; Pietzner; 2014) and its assessment (Osman & Kaur, 2014). Computers have aided in teaching and learning using various forms. For example, a teacher may use the device to show how a reaction progresses in Chemistry (Meijer & Nering, 1999; Pietzner; 2014). Further, when the computer is connected to the internet, it can be used as a medium for forum discussions both in and out of school (Ozyurt & Ozyurt, 2013). Similarly, the computer has been used as an essential assessment aid from the time of its invention. One of its important assessment tools is computerized adaptive testing.

Computerized adaptive testing (CAT) is a computer-based test that has been programmed to tailor to the test takers' ability (Wainer, 2000; Reckase, 2009). It has

Correspondence: Bor-Chen Kuo,
Graduate Institute of Educational Information and Measurement, National Taichung University of Education, Minsheng Rd., West Dist., Taichung City 40306, Taiwan.
E-mail: kbc@mail.ntcu.edu.tw
doi: 10.12973/eurasia.2015.1384a

ISSN: 1305-8223

two types; unidimensional computerized adaptive testing (UCAT) and multidimensional computerized adaptive testing (MCAT). UCAT conducts an overall ability estimate of the test takers, and MCAT looks beyond that for sub-score information across different tested domains (Wainer, 2000; Reckase, 2009). UCAT was first used in educational assessment as American university placement tests in 1994 (Meijer & Nering, 1999). While, UCAT development in Indonesia has been started by Kustiyahningsih and Cahyani (2013) for the subject of the Indonesian language. According to Reckase (2009), after 1999 most available CATs are MCATs. Wainer (2000) explains that in addition to information regarding sub-scores, MCAT provides correlation estimations for item responses between domains. This information improves the measurement precision of sub-scores compared to UCAT. Examples of MCATs are the Math MCAT in the Netherlands and South Korea (Meijer & Nering, 1999; Chae, Kang, Jeon, & Linacre, 2000) and the Chinese CAT in Taiwan (Wang, Kuo, & Chao, 2011; Wang, Kuo, Tsai, & Liao, 2012).

This study set out to develop a Biology MCAT for Indonesian junior high school students. The reason for developing MCAT and not UCAT is because Indonesian curriculum as national standard has 6 Biology domains (The Ministry of National Education, 2006). So, it is by nature the curriculum goes aligned with MIRT which consequently goes in line with MCAT for its learning assessment. In addition to that, obtaining more information in different Biology domains of the test will be a good resource for remedial instruction.

This effort is perceived as an innovation to the Indonesian National Examination (NE). The NE has used a paper and pencil (P&P) test mode for more than three decades. Its purpose is as a tool to improve the quality of education in Indonesia (Firman & Tola, 2008). However, measurement precision and leaking of items have become two recent issues (Sulistyo, 2009; Rahmi, 2011; Maryono & Purnama, 2012; Solopos, 2014; Kuo & Daud, 2014b; Kuo & Daud, 2015).

State of the literature

- Computerized adaptive testing (CAT) has better measurement precision comparing to paper and pencil (P&P) test mode. It also has better item security protection so that leaking of the test material, as was experienced by the Indonesian National Examination (NE), can be avoided.
- Multidimensional computerized adaptive testing (MCAT) is a newer version of CAT. It looks the sub scores of test takers from the different domains tested. In addition, MCAT measurement precision is also higher than that for UCAT because it considers correlation estimation among its domains.
- Evaluation of MCAT with different methods determines the best performance of the system. The process also improves measurement precision with less bias and error of the system.

Contribution of this paper to the literature

- This study is the first MCAT development in an Indonesian context. The design, method, and process were constructed based on the Indonesian Biology curriculum. Our findings lay the foundation for future study.
- It uses empirical data to evaluate MCAT performance so that the result of the evaluation is more reliable than one that might be based on generated data. Root-mean-squared error (RMSE) method is used. The bias and standard error (SE) of the system are reported to assure the system is workable and consistent in its measurement precision, with a limited number of the items administered compared to the P&P test.
- This MCAT demonstrates how measurement precision and its item security protection can be gained, along with a comparison to the NE P&P test mode.

Multidimensional item response theory

Multidimensional item response theory (MIRT) is an extended model of unidimensional item response theory (UIRT). One important MIRT model is the multidimensional random coefficients multinomial logit (MRCML) model that was suggested by Adams, Wilson, and Wang (1997). With the MRCML, an item is viewed as one of two types: a between-item multidimensional test or a within-item multidimensional test. According to Wang et al. (2011), the difference between these two types of tests is the domain(s) of one item that is measured during the test.

Every item in a between-item multidimensional test is measured in one domain only. On the contrary, for a within-item multidimensional test, each item is measured in more than one domain simultaneously (Wilson & Adams, 1995; Adams et al., 1997) using the following formula:

$$P(\mathbf{X}_{ik}=1;\mathbf{A},\mathbf{B},\xi\,|\,\boldsymbol{\theta})=\frac{\exp(\mathbf{b}_{ik}^{'}\boldsymbol{\theta}+\mathbf{a}_{ik}^{'}\xi)}{\sum_{k=1}^{K_i}\exp(\mathbf{b}_{ik}^{'}\boldsymbol{\theta}+\mathbf{a}_{ik}^{'}\xi)}$$

where X_{ik} is a response of category k in item i; K_i is the number of categories of the item i; $\boldsymbol{\theta}$ stands for the participant's multidimensional ability which are represented by the vector $\boldsymbol{\theta}=(\theta_1,\theta_2,\ldots,\theta_D)'$; ξ indicates the vector of the item parameters. A scoring matrix, $\mathbf{B}$, that represents the relationships between items and dimensions; $\mathbf{b}_{ik}=(b_{ik1},b_{ik2},\ldots,b_{ikD})'$ represents the scores for category k across D dimensions, then composes to $\mathbf{B}_i=(\mathbf{b}_{i1},\mathbf{b}_{i2},\ldots,\mathbf{b}_{iD})'$ for item i, and then composes to $\mathbf{B}=(\mathbf{B}_1^{'},\mathbf{B}_2^{'},\ldots,\mathbf{B}_n^{'})'$ for the entire test. $\mathbf{A}=(a_{11},a_{12},\ldots a_{1K_1},a_{21},\ldots a_{2K_2},\ldots,a_{nK_n})$ is called design matrix that represents the relationships between items and the model parameters, such as the item difficulties.

MCAT procedures

There are four steps to set up a CAT: initialization of how to start the test; item selection rules of how to administer items to test takers; ability estimation of how to predict proficiency of test takers; and setting a stopping rule.

In the initialization step, the test taker is estimated to have a score of 0 in its initialization in most CAT practices (Thomson & Weiss, 2012). Thomson and Weiss (2012) suggest that a randomization technique is used, such as considering a scale from -0.5 to +0.5 of item difficulty as initialization. Another suggestion is to consider the background information of students that can be gathered from classroom tests or even pre-tests conducted prior to CAT initialization (Kustiyahningsih & Cahyani, 2013). To some extent, the initial ability estimation in MCAT uses one of these existing procedures where the initial ability (θ) estimate is assumed to be 0. This initial ability estimation also has been used in MCAT systems, such as Chinese proficiency test (Wang et al., 2011). Its mean of the prior distribution of ability μ = (0, 0) and standard deviation (SD) set is 1 (Segall, 1996; Wainer, 2000; van der Linden & Glas, 2002).

In the item selection procedure, Segal (1996) suggests that items can be administered on the basis of item information. This provisional ability estimate obtained from the response of the j-th item is used to evaluate the item information function. The discourse of item selection continues where van der Linden and Glas (2002) suggest the item selection procedure should rely on a maximum information criterion and Owen's Approximate Bayes Procedure. With this procedure, Reckase (2009) adds, CAT estimates the location of the examinee's ability on a coordinate system. However, Segall (1996) suggests that Maximum Likelihood Estimation (MLE) is used to produce a conditional distribution of ability estimates in combination with the Fisher information matrix. According to Wang et al. (2011), two methods widely used in MCAT are maximum information and maximum expected precision information. Based on the information gathered by these methods, the next item is administered to the test taker.

For ability estimation, maximum likelihood estimation (MLE) has been used in CAT (van der Linden & Glas, 2002). However, it has a weakness in that it requires a mixture pattern of right and wrong answers together to get a precise estimation (Thomson & Weiss, 2012). As a result, joining MLE with Bayesian estimation of expected a posterior (EAP) is used, especially when only a small number of items is administered. Maximum a posterior (MAP) estimation is much more appropriate for a large number of items (Reckase, 2009; Babcock & Weiss, 2012). Chen (2006) adds that MAP is appropriate for MCAT systems that test a larger number of domains. MCAT also adopts this existing procedure for MLE, EAP, and MAP of Bayesian procedure for its ability estimation. An example of MCAT that uses a MAP estimator is MCAT for Chinese proficiency test (Wang et al., 2011).

For the stopping rule, there are two types: fixed test length and varying test length. The fixed test length is the specified number of test items to be administered to the test taker (Reckase, 2009; Yao, 2012). In contrast, the varying test length is an estimated ability of when the desired precision level or confidence level has been reached (Wainer, 2000; Yao, 2012). In addition, Chao, et al. (2000) also addressed another CAT stopping rule; when the item bank is exhausted and the ability measure is far enough away from the pass-fail criterion, and the test taker is exhibiting off-test behavior. Early CAT application prefers varying test lengths because it is in line with the intension to produce better precision for test taker ability estimates (Wainer, 2000; Yao, 2012; Babcock & Weiss, 2012). However, for some practices, this would prolong the test time because the test taker would be assumed to be a smart test taker. He/she would even have an effect on item exposure since during a longer test, the test bank would have fewer items left over. As a result, scholars started to use a fixed test length as a stopping rule (Yao, 2012).

MCAT scales and item bank development

The scale refers to a common measurement tool that can estimate all abilities from different backgrounds with a valid measure of an underlying construct (Clark &Watson, 1995; DeVellis, 2003, Nguyen, 2010). Wainer (2000) and Thomson and Weiss (2012) suggest that items, in scales, can be used as an assessment tool as they are stored in the CAT item pool/item bank. However, they have to be pre-tested and parameterized in advance. Therefore, developing a scale and item bank can be done concurrently.

Scale development starts with setting a goal to determine what kind of scale it will be and how it will be developed (Clark & Watson, 1995; Hinkin, 1995). For example, a scale is to be developed to measure students' abilities in math for grade 6. In this case, the math curriculum along with related learning materials and guidelines are used. The next step is to design an item blue print. This guide ensures content balance of topics for particular subjects to be tested (Clark & Watson, 1995). Data collection also needs to be balanced so that samples with different backgrounds have an equal opportunity to take part in the study (Kuo & Daud, 2014a).

According to Thomson and Weiss (2012), a bank with 400 to 500 items would be of high quality. It would allow the other 100 items to be rotated into different positions when administered to test takers. Some CATs also use fewer items; Veldkamp and van der Linden employed a 176-item pool (as cited in Reckase, 2009), while Kustiyahningsih and Cahyani (2013) had 180 items in their CAT item bank. Reckase (2009), states that "with an MIRT CAT of 60 items, it seems to be enough information to overcome the influence of the prior distribution so the bias seems to be minimal" (p. 324). However, the purpose of CAT is to obtain better precision of test taker ability, which means it would be better if more items were available (Babcock & Weiss, 2012). Considering its time limitations and financial constraints, this study developed 300 items for its MCAT item bank.

Evaluation of MCAT performance

The performance of a CAT must be evaluated at some point to determine if the system functions effectively. Most CAT evaluation focuses on the procedures of ability estimation precision, item selection procedure and stopping rule (Wainer, 2000; Reckase, 2009; Wang, et al., 2011; Yao, 2012; Wang, Chang, & Boughton, 2012). This evaluation is to assure the accuracy of the system in its measuring ability estimation as well as to reduce any errors in the precision. To achieve this outcome, some scholars use generated data, while others use real data. At the same time, some of them use both real data and generated data to observe the different performances with difference simulations. This method allows scholars to condition the system with various situations.

For ability estimation, different methods have been used by scholars; these methods include maximum likelihood estimation (MLE), expected a posterior (EAP) estimation, weighted likelihood estimation (WLE), and MAP (Wainer, 2000; Reckase, 2009). Besides, root mean squared error (RMSE), bias and standard error (SE) are three important indexes they look at to guarantee the system has less error (Wainer, 2000; Reckase, 2009). Wang, et al. (2011) used the methods of MLE, EAP, and MAP to determine how their MCAT ability estimation performed. Their result indicates that MAP is better for their system than MLE and EAP. Meanwhile, some scholars also look at bias and SE indices. Babcock and Weiss (2012) suggest that the criterion of bias is 0.16, while the SE falls between 0.385 ~ 0.315. Lower bias criteria were suggested by Wang, Hanson, and Lau (1999) to consider any bias value is below 0.01. SE is used to assess the performance of all item information and fluctuations across administered items to test takers. These formulas are shown below.

$$RMSE(\hat{\theta}_i) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)^2}$$

$$Bias(\hat{\theta}_i) = \sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)$$

$$SE(\hat{\theta}_i) = \frac{1}{\sqrt{I(\hat{\theta}_i)}}$$

where θ_i represent the i^{th} participant's ability which is estimated from the ConQuest, $\hat{\theta}_i$ represents the i^{th} participant's temporarily ability estimation after partial items had been responded; N represents the total number of participants, and $I(\hat{\theta}_i)$ is item information.

For the item selection procedure and stopping rule, Yao (2012) evaluated her MCAT using generated data to compare the performance of five different simulations. She found that Kullback-Leibler's (KL) information is better for varying the length of her MCAT; Wang et al. (2012) based their evaluation on both empirical data and generated data for their Math MCAT's stopping rule. Their finding was that a maximum Kullback-Leibler divergence rule was better for their MCAT.

METHODOLOGY

Participants

Participants of this research were 2238 (originally 2314) grade 9 students from 24 junior high schools, between the ages of 13 and 14 years old. About 40% of these were boys (1019), and 60% of them were girls (1219). To further classifying gender according to types of schools, there were 738 boys and 933 girls at public schools and 281 boys and 286 girls in private schools. While looking into location, it was found that there were 392 boys and 452 girls who attended rural schools of four districts and 627 boys and 767 girls at urban schools. All participants of the study were from the Aceh province and might not represent the entire country of Indonesia. That is limitation of this study.

Instrumentation

Item construction was carried out together with a team of Biology teachers. Before the items were tested on the students, they were reviewed by experts from a local university in Aceh Indonesia. Based on the Indonesian Biology curriculum, the developed item blueprints provide framework in item construction.

There are 6 domains; 1) Biology and Research; 2) Botany; 3) Zoology; 4) Human beings; 5) Anatomy Function and; 6) Ecosystem). Every domain has 3 to 5 sub-domains. Domain 1 has 3 sub domains. Domains 2, 3, and 4 have 4 sub-domains each. Meanwhile, domains 5 and 6 are comprised of 5 sub-domains each. In total, there were 25 sub-domains. In every sub-domain, 12 items are constructed, which is summed up in total up to 300 items.

ConQuest (Wu, 2012) was used for item calibration. The mean squared (MNSQ) criteria used were from 0.8 to 1.20. This is in line with consideration of programs for international student assessment (PISA) criteria (OECD, 2012).

Biology MCAT procedures

Figure 1 shows MCAT procedures. In its initialization, test taker's ability (θ) is set as 0. While in item selection, the system will select an item from the bank with the maximum item information on current estimated ability. After administering the selected item, MCAT will estimate examinee's ability based all responses of administered items. Then MCAT will check the stopping rules are satisfied or not. If one of the stopping rules is satisfied, then system will stop giving items and report the examinee's proficiencies. If none of the stopping rules is satisfied, then MCAT will do item selection again based on the new obtained ability. Usually, there are three types of the stopping rules. First is the criterion of SE, second is the test length, and third is the test time (Wainer, 2000; van der Linden and Glas, 2002; Wang, et al., 2012)

In this study, a simulation study based on the item and ability parameters estimated from real data is conducted to evaluate the performance of the proposed

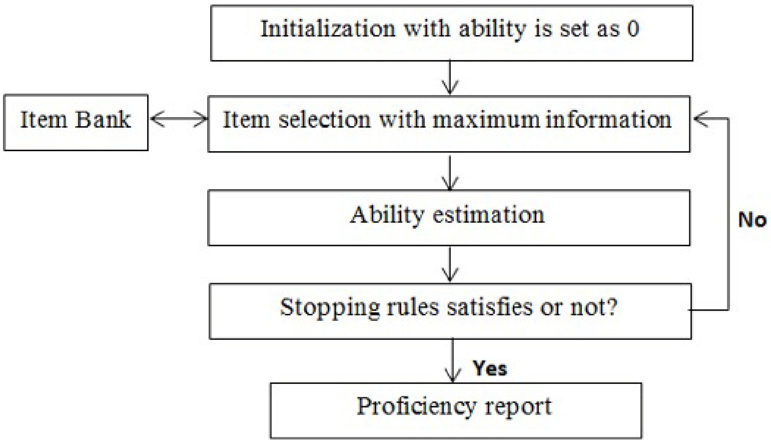
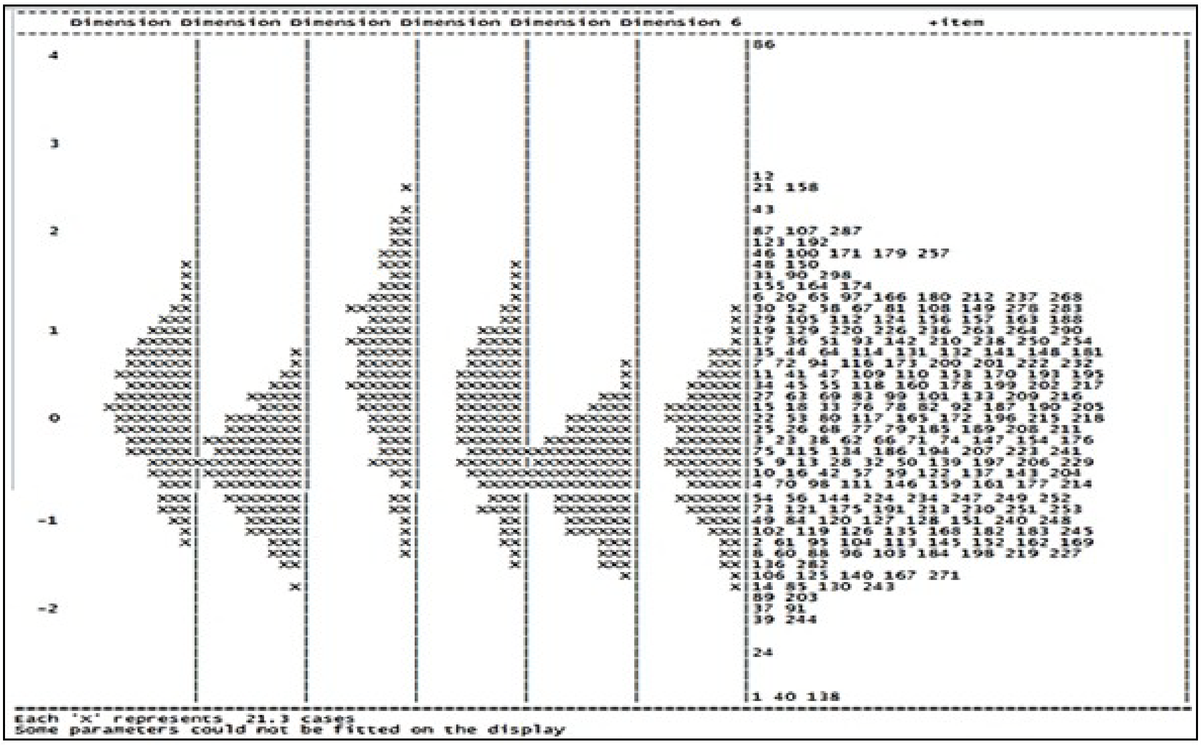

Initialization with ability is set as 0
Item Bank
Item selection with maximum information
Ability estimation
No
Stopping rules satisfies or not?
Yes
Proficiency report

Figure 1. MCAT Procedures



```
     Dimension Dimension Dimension Dimension Dimension Dimension 6          +item
-----------------------------------------------------------------------------------------
 4                                                               |86
 3
                                                                 |12
                                                                 |21 158
                                                                 |43
 2                                                               |87 107 287
                                                                 |123 192
                                                                 |46 100 171 179 257
                                                                 |48 150
                                                                 |31 90 298
                                                                 |155 164 174
                                                                 |6 20 65 97 166 180 212 237 268
                                                                 |30 52 58 67 81 108 149 278 283
                                                                 |29 105 112 124 156 157 163 188
 1                                                               |19 129 220 226 236 261 264 290
                                                                 |17 36 51 93 142 210 238 250 254
                                                                 |35 44 64 114 131 132 141 148 181
                                                                 |7 72 94 116 173 200 201 222 232
                                                                 |11 41 47 109 110 153 170 193 195
                                                                 |34 45 55 118 160 178 199 202 217
                                                                 |27 63 69 83 99 101 133 209 216
 0                                                               |15 18 33 76 78 82 92 187 190 205
                                                                 |22 53 80 117 165 172 196 215 218
                                                                 |25 26 68 77 79 185 189 208 211
                                                                 |3 23 38 62 66 71 74 147 154 176
                                                                 |75 115 134 186 194 207 221 241
                                                                 |5 9 13 28 32 50 139 197 206 229
                                                                 |10 16 42 57 59 122 137 143 204
                                                                 |4 70 98 111 146 159 161 177 214
                                                                 |54 56 144 224 234 247 249 252
                                                                 |73 121 175 191 213 230 251 253
-1                                                               |49 84 120 127 128 151 240 248
                                                                 |102 119 126 135 168 182 183 245
                                                                 |2 61 95 104 113 145 152 162 169
                                                                 |8 60 88 96 103 184 198 219 227
                                                                 |136 282
                                                                 |106 125 140 167 271
                                                                 |14 85 130 243
-2                                                               |89 203
                                                                 |37 91
                                                                 |39 244
                                                                 |24
                                                                 |1 40 138
=========================================================================================
Each 'X' represents   21.3 cases
Some parameters could not be fitted on the display
```

Figure 2. Six Biology domain scales

Biology MCAT system with two ability estimation methods, MLE and MAP. Three indexes, SE, bias, and RMSE of ability estimations, are used for the evaluation. After evaluation, the ability with better performance will be implemented in Biology MCAT system.

RESULT

Biology scale and item bank

The first result of the study was the Biology scales and the parameterized item bank. The scales is integrated into the MCAT system and used as plat-forms to measure the test takers' competencies against these scales.

Figure 2 shows the Biology domain scales; 1) Biology and Research scale; 2) Botany scale; 3) Zoology scale; 4) Human Being scale; 5) Anatomy Function scale; and 6) Ecosystem scale. It can be seen that the domain 3 (Zoology) scale contains the most difficult items, as it requires a score of about 2.6 for the ability value to answer the most difficult items correctly. In contrast, the domain 2 (Botany) scale has the easiest items where test taker needs to have a score of -1.8 to get its items correctly. Based on this distribution, three patterns emerge where the item distribution for the domain 2 (Botany) scale and the domain 4 (Human Being) scale is much more concentrated, and the range was between -1.8 to 0.8. In general, the distribution of items with these six scales is reasonably normal.

It is reported that based on criteria set for model fit (<0.8 to >0.2), four items were out of criteria. Those items are item 91, 124, 127 and item 128. These items have been deleted from scale, and as a result there is 296 items left in item bank. While, reliability coefficient Cronbach's alpha of each domain is as follows: 0.846 for domain 1; 0.807 for domain 2; 0.829 for domain 3; 0.866 for domain 4; 0.807 for domain 5; and 0.852 for domain 6.

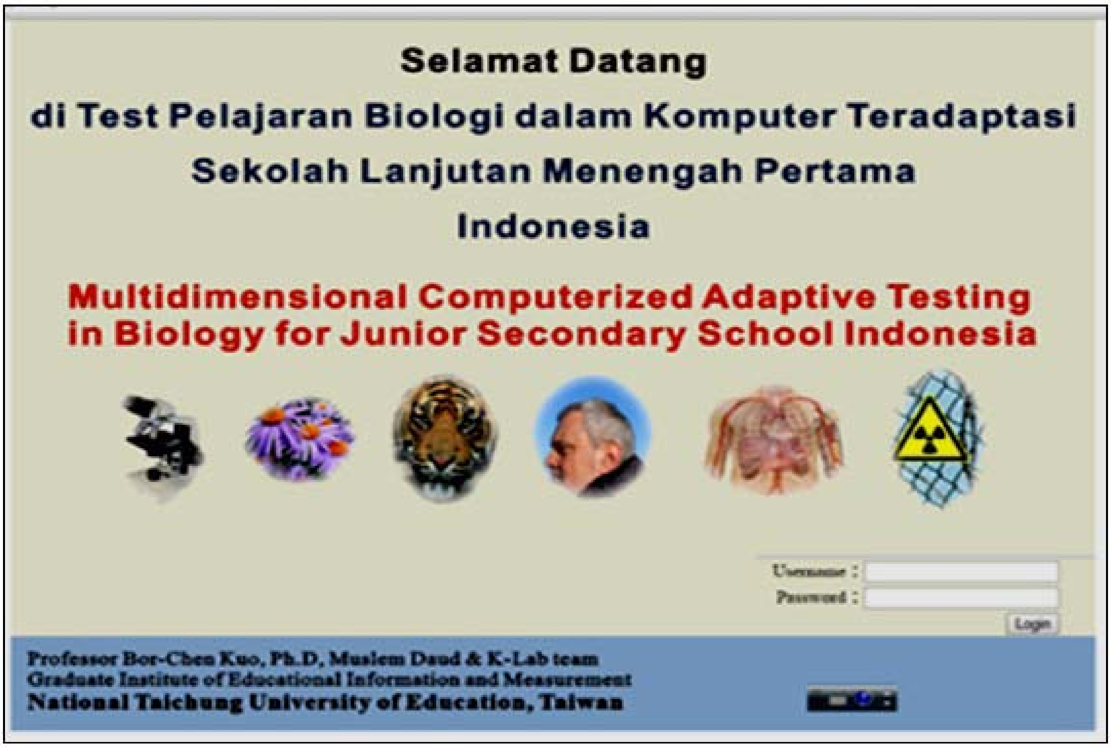
Biology MCAT system

The second result of the study is the Biology MCAT system. The MCAT is integrated to a server and can be accessed online (http://210.240.193.249/irtm/). This HTML system is connected to a Biology item bank. A test taker needs an identity (ID) and a password to log in and sit for the test. This requirement is part of the server security. The following figures explain interface of webpage, example of item and report card provided for test taker at the end of the test.

Figure 3 shows the webpage's greeting in the Indonesian language, which explains that the webpage is an adaptive testing tool in a multidimensional Biology domain for Indonesian junior high school. The multi-dimensionality image is represented by pictures of laboratory tools, plantations, animals, humans, human anatomy, and ecosystem protection signs. ID and password columns are provided at the corner of the page and need to be completed if the test taker is to sit for the test.

During initialization, this system set the ability estimate (θ) at 0. The system chose the first item in the middle of a range from very easy to very difficult from all six dimensions. Then, the system administered the first item. For the next item, the system picked an item from the item bank with a difficulty parameter close to the test taker's current ability estimation. This behind-the-scenes process of updating based on test taker ability could get progressively easier or more difficult or even stabilize for some time depending on the response given. Even though two test takers might sit next to each other while taking the test, their items will not be the same because their ability estimate would be different. However, in this process, students who do well will get more difficult items compared to those whose competency is average.

Figure 4 shows example of Biology MCAT item. The test asks test taker to look at the picture in order to response to the question. He/she then needs to choose one best answer by pressing one small dot in front of every alternative answers. After that, he/she to press enter key on computer keyboard to get the next item.



Selamat Datang

di Test Pelajaran Biologi dalam Komputer Teradaptasi

Sekolah Lanjutan Menengah Pertama

Indonesia

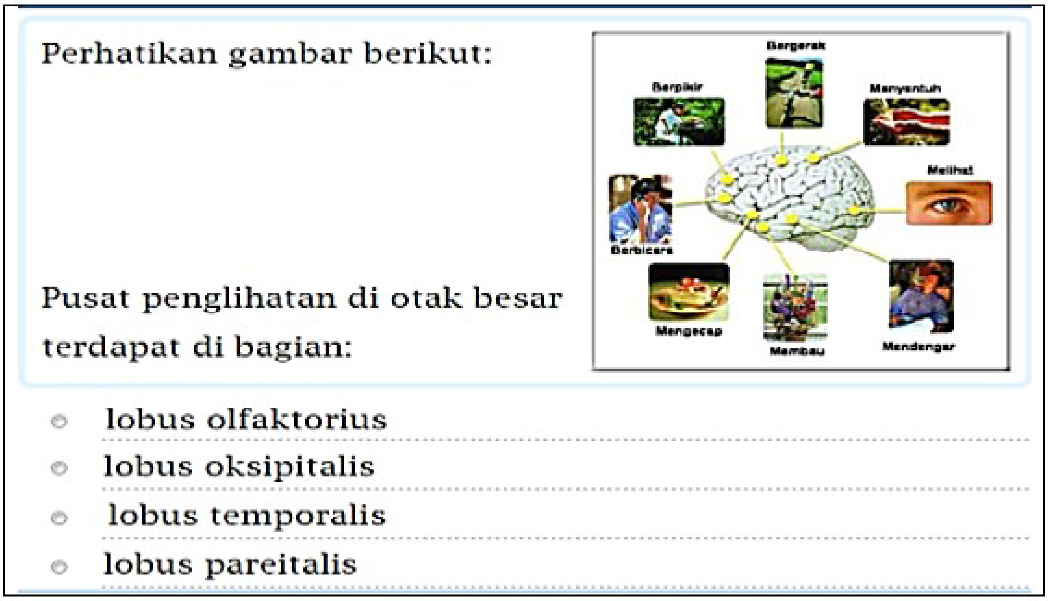
Multidimensional Computerized Adaptive Testing
in Biology for Junior Secondary School Indonesia

Username :

Password :

Professor Bor-Chen Kuo, Ph.D, Muslem Daud & K-Lab team
Graduate Institute of Educational Information and Measurement
National Taichung University of Education, Taiwan

Figure 3. MCAT webpage

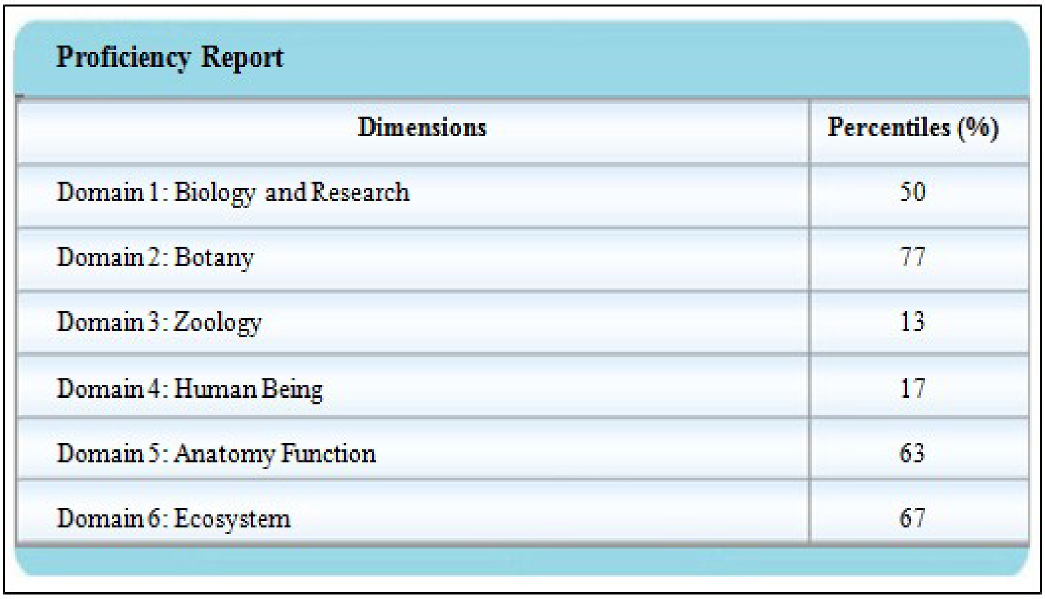


Perhatikan gambar berikut:

Pusat penglihatan di otak besar terdapat di bagian:

- ○ lobus olfaktorius
- ○ lobus oksipitalis
- ○ lobus temporalis
- ○ lobus pareitalis

Figure 4. Example of MCAT item



Proficiency Report

Dimensions	Percentiles (%)
Domain 1: Biology and Research	50
Domain 2: Botany	77
Domain 3: Zoology	13
Domain 4: Human Being	17
Domain 5: Anatomy Function	63
Domain 6: Ecosystem	67

Figure 5. MCAT proficiency report

At the end of test, every test taker will be provided with report card. The report tells test taker's performance on the test. Part of report card is proficiency report as shown in Figure 5. Figure 5 shows proficiency information across 6 multi-dimensions of test taker's proficiency toward subject tested. Test-taker got a highest achievement in domain 2 (Botany) and the lowest in domain 3 (Zoology). Another lowest achievement is in domain 4 (Human Being). While achievement in other domains; domain 1 (Biology and Research), domain 5 (Anatomy Function) and domain 6 (Ecosystem) are ranged from 50% to 67%.

The evaluation results of Biology MCAT system

In this section, the RMSE, SE, and bias of MLE and MAP are reported and compared. Figure 6 shows RMSEs of MLE and RMSE of MAP across six Biology domains. On overall, RMSEs of MLE shows fluctuation lines for the early trends of all domains. They gradually decrease and stabilize when more items have been

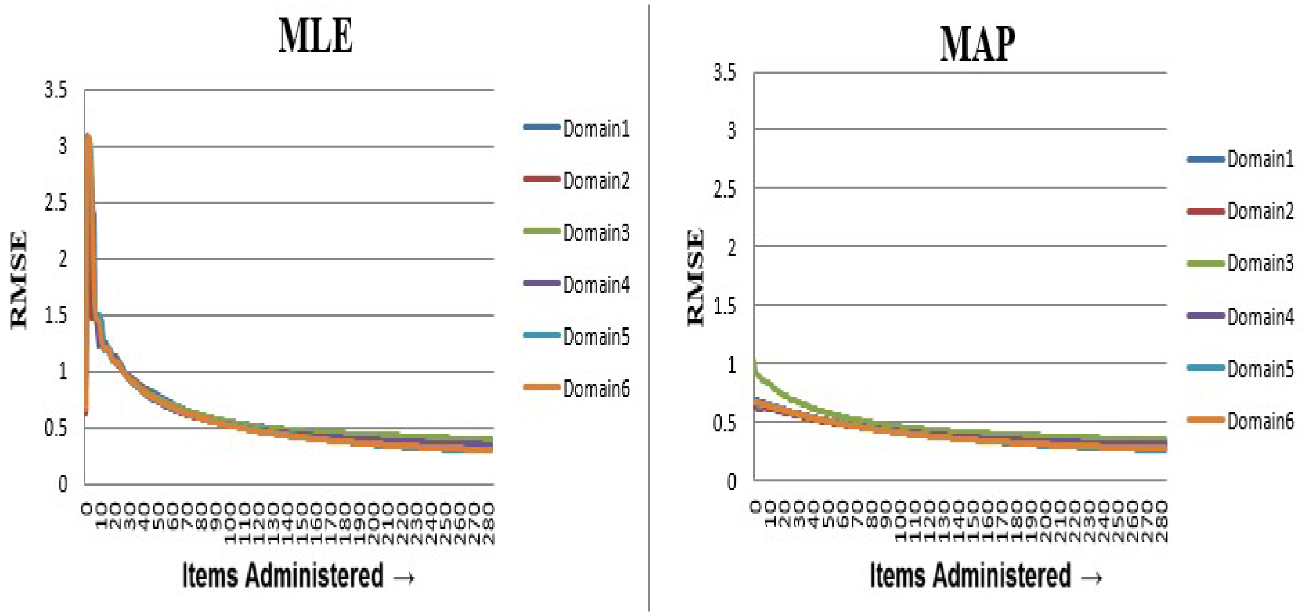


Figure 6. The RMSEs of MLE and The RMSEs of MAP estimation among the six domains

answered. It started at 0.6, jumped to 3.05 at the second item and declined to 1.5 at the third item. The instability continues for the following items until test takers had responded to the 16th item where the RMSE of receptive activities was 1.3. The lines became even more stable following items 31 (0.9) and 46 (0.8).For item 61, the receptive activities and strategies for the six domains was about 0.6. The variation of the MLE estimation among the six domains steadily decreased until all items had been answered.

In the meantime, RMSEs of MAP casts steadily decreasing lines, starting with an early trend and continuing until the end of all items have been answered. The MAP figure shows that when test takers responded to the 16th item, the RMSE of receptive activities was 0.6. The lines become more stable after the 31th (0.57) and 46th (0.52) item. For item 61, the receptive activities and strategies domain number was 0.48. Specifically for domain 3, the line started at 1.0 and stayed away from other domains' lines until item 61. Compared to RMSEs of MLE to RMSEs of MAP where domain 2 is much more fluctuated, RMSE of MAP in domain 2 is more stable in its early stage. Overall, MAP estimation process worked better with the system than that in RMSEs of MLE.

Figure 7 shows MLE and MAP Bias. MLE bias trend shows a fluctuation range among the six domains in the early stages of the test where the estimated ability is far away from the actual ability (-0.71 ~0.84). The bias of domain 1 indicates the biggest difference than other domains, especially domain 6. The estimation trend of all domains is moving closer together at the 7th item: 1.06. At the 16th item, the bias is 0.14, while it was 0.02 for item 31. At the 46th item, the bias is 0.03, and it is also 0.03 for item 61. At this point, the bias stabilizes between 0.02 and 0.03 until the items end. According to Babcock and Weiss (2012), 0.16 is an acceptable value for the bias; therefore, a test taker's ability estimation becomes precise after administering 15 items (0.15).

The figure also shows the wide range of the MAP bias trend among the six domains in the early stages of testing where the estimated ability is far away from the actual ability of 1.06. The bias of domain 3 indicates the biggest difference among all the domains. The estimation trend of all domains becomes more similar at item 7 (0.83). At item 16, the bias is 0.6, while it decreases to 0.5 for the 31st item. At item 46, the bias is 0.4, and it decreases to 0.3 at item 61. The bias becomes smaller and smaller until the end of items is reached at a bias of 0.12. Based upon Babcock and Weiss's (2012) acceptable value of 0.16 for bias, these findings indicate that MLE bias performs better than the MAP bias.

Figure 8 shows standard error of MLE and MAP estimation. It shows that SE of MLE trend starts with the first item's large margin of error and then dramatically descends to a lower error by the 3rd item. The line then steadily decreases until item 16th, where its value is 3.11 of error rate. Then, this SE index considerably stabilizes in its decreasing mode, where it is 0.519 at item 31. The error rate goes below 0.3 starting from item 35 where it becomes 0.370, and 0.168 for item 46. After that, the index goes flatly stable until the end of all items (0.01). Babcock and Weiss (2012) reported that values from 0.385 ~ 0.315 are acceptable SE criteria; therefore, the test can be concluded after administering 35 items (0.370).

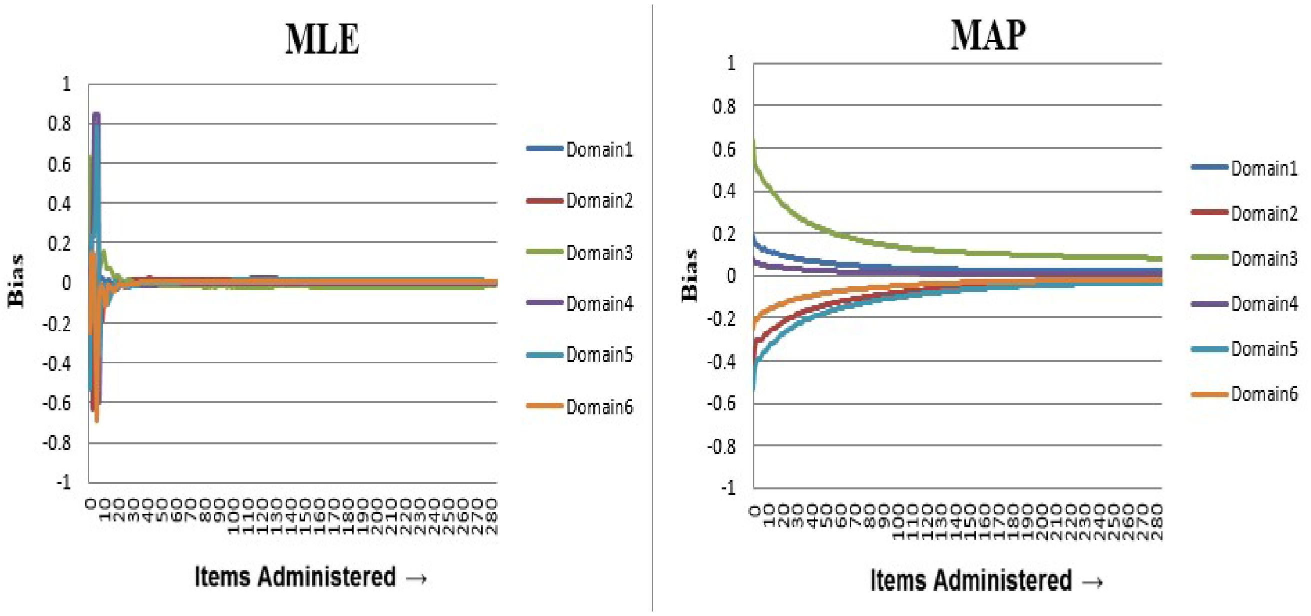


Figure 7. The bias of MLE and MAP estimation

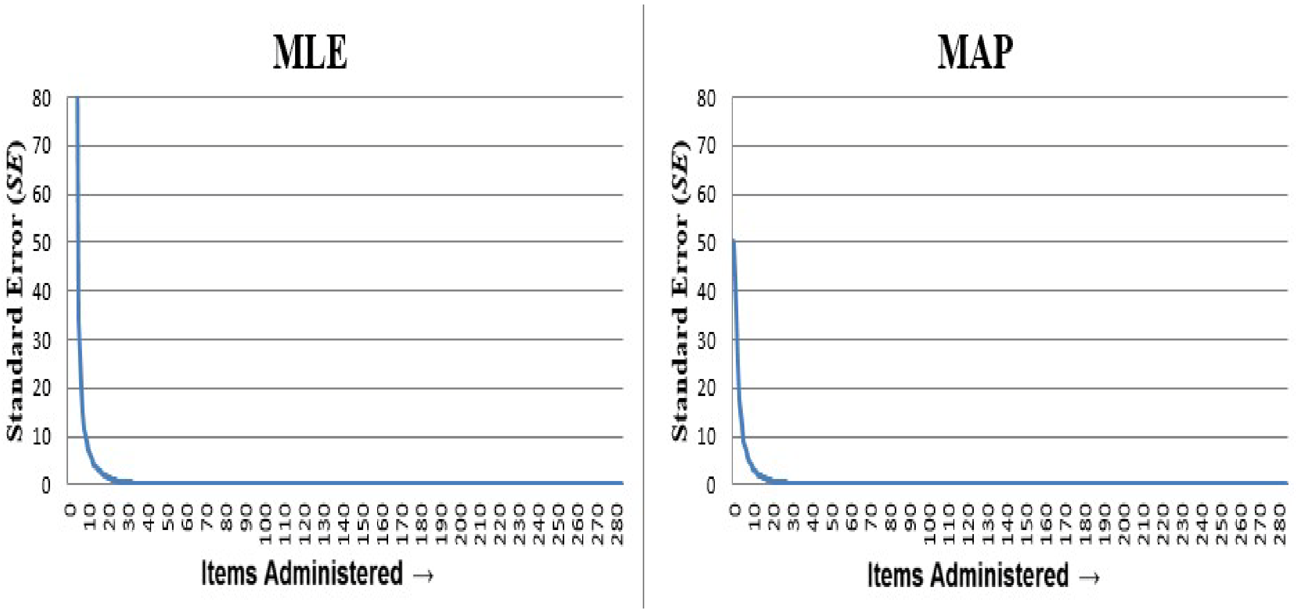


Figure 8. The standard error of MLE and MAP estimation

The figure also shows that SE of MAP trend starts with a large error of 50 and then dramatically decreases to 7.368 at the 7th item. The line then further declines until item 16, where its value is 1.515. Then, SE index considerably stabilizes in its decreasing mode. The error rate goes below 0.3 starting from item 29 where it becomes 0.372, and 0.315 at item 31. It decreases further to 0.111 for item 46. This value goes flatly stable until the end of all items (0.01).

According to Babcock and Weiss (2012), 0.385 ~ 0.315 is an acceptable SE. Therefore, the test can be concluded at item 29 (0.372). Compared to the SE of MLE estimations, the SE of MAP shows less error with fewer items administered.

Based on the results of the SE of MAP evaluation, the system could stop after administering 29 items (0.372). The values shown indicate an acceptable error rate for the system.

DISCUSSION

This study has constructed the MCAT system in Biology with six domains in Indonesian context. The evaluation result shows that MAP ability estimation has better performance than MLE estimation, and this finding is similar to Chen (2006). Figures 7 also points out that the bias value is below 0.16 after only limited items have been administered. This means the constructed Biology MCAT has small measurement error when a participant takes the Biology test. When acceptable SE criteria reach from 0.315 ~ 0.385 (Babcock &Weiss, 2012), the test can be ended after administering 35 items (0.370) of MLE SE (Figure 8). However, the test could be stopped even earlier after administering 29 if MAP SE is used because it reaches the acceptable criteria earlier than the MLE SE.

MCAT system is more convenient to hold examination than to deliver the P&P test to difference parts in Indonesia. MCAT system also minimizes the leaking of test materials, which is commonly experienced by traditional P&P tests, such as the Indonesian NE (Kuo & Daud, 2014b; Kuo & Daud, 2015). The leaking could take place when test materials (items) are printed, photocopied, or transported to or from the central test center for grading. Furthermore, MCAT system's adaptive function is set to follow the progression of individual test taker ability. In other words, all test takers will have different items even though they may access the test at the same time and from the same place. So, the chance of cheating is lower, and therefore the test will be more reliable.

Although it brings a lot of benefit, there are issues have not been covered in this study, such as content balance and item exposure rates. The item selection of maximum information method in this MCAT system is only considered the item information to temporally ability estimation. That means some participants may take a lot of items in the same domain or lack items from some domains.

While for item exposure rates, the maximum information method will always select the item which has the highest information without any constraints. For example, all participants will get the same item which has the highest information in the beginning of the adaptive test. This situation will lead to some items with high information are not used and therefore it shortens the life of item bank. These two issues shall be concerned to make the MCAT system more comprehensive to future applications.

CONCLUDING COMMENTS

In this study, an MCAT system with six Biology domains are developed based on MRCML model and a real data is used to evaluate its performance. These six Biology domains includes 1) Biology and Research; 2) Botany; 3) Zoology; 4) Human Being; 5) Anatomy Function; and 6) Ecosystem domains.

The results indicate that it functions well in many facets: 1) the measurement error are small enough to acceptable, 2) participants take fewer items to finish the test when the system selected items adaptively, 3) and the system can give feedback to participants immediately. The domain scores can provide the information about the strengths and weaknesses of test takers and this information would be helpful for remedial instruction.

Finally, MCAT implementation also has implied its simplicity; it can make NE much simpler regarding logistical issues by considering Indonesia's wide demographic area and large number of NE participants. In this case, the MCAT online system would be more reliable compared to NE P&P tests. Thus, the expectation to convert the NE to a tool that can improve the quality of education Indonesia can be achieved.

REFERENCES

Adams, R. J., Wilson, M. R., & Wang, W.C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1–23.

Babcock, B., & Weiss, D. J. (2012). Termination criteria in computerized adaptive tests: Do variable-length CATs provide efficient and effective measurement? *Journal of Computerized Adaptive Testing, 1*(1), 1–17.doi: 10.7333/1212-0101001

Chae, S., Kang, U., Jeon, E., & Linacre, J.M. (2000). *Development of computerized middle school achievement test.* Seoul: Komesa Press.

Chen, P. H. (2006). The influences of the ability estimation methods on the measurement accuracy in multidimensional computerized adaptive testing. *Bulletin of Educational Psychology, 38*(2), 195–211.

Clark, L.A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*(3), 309–319.

DeVellis, R. F. (2003).*Scale development: Theory and applications*, 2nd ed., London: SAGE Publications.

Firman, H., & Tola, B. (2008). The future of schooling in Indonesia. *Journal of International Cooperation in Education*, *11*(1), 71–84.

Hinkin, T. R. (1995).A review of scale development practices in the study of organizations. *Journal of Management*, *21*(5), 967–988.

Kuo, B. C., & Daud, M. (2014a), Biology Scale Development for junior secondary school students, Aceh Indonesia. In *The fifth Pacific Rim Conference on education: Educational innovation* (pp.109–110). Taipei: University of Taipei.

Kuo, B. C., & Daud, M. (2014b). Computerized adaptive testing as an innovation for Indonesian national examination. In *International Multidisciplinary Conference* (pp.223-231). Jakarta: Jakarta Muhammadiyah University.

Kuo, B. C., & Daud, M. (2015). *CAT untuk UN jujur* [CAT for an honest national examination]. Retrieved April 14, 2015, http://aceh.tribunnews.com/2015/04/14/cat-untuk-un-jujur

Kustiyahningsih, Y., & Cahyani, A. D. (2013). Computerized adaptive test based on item response theory in e-learning system. *International Journal of Computer Applications, 81*(6), 6-11.doi: 10.5120/14014-2022

Maryono, M., & Purnama, B.E. (2012). Education policy development with development strategy application of national test exercises for vocational high school: Case study vocational high school Bina Taruna Masaran Sragen. *International Journal of Computer Science* Issues, *9*(5), 136–145.

Meijer, R. R., & Nering, M.L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement, 23*(187).doi:10.1177/01466219922031310

Nguyen, P.C. (2010). Author guidelines for reporting scale development and validation results in the Journal of the Society for Social Work and Research. *Journal of the Society for Social Work and Research, (1)*2, 99–103.

OECD. (2012). *PISA 2009 technical report PISA*. OECD Publishing. doi: http://dx.doi.org/10.1787/9789264167872-en

Osman, K., & Kaur, S. J. (2014). Evaluating biology achievement scores in an ICT integrated PBL environment. *Eurasia Journal of Mathematics, Science & Technology Education, 10*(3), 185–194.

Ozyurt, O., & Ozyurt, H. (2013). An examination of computer engineering students' perceptions about asynchronous discussion forums. *Eurasia Journal of Mathematics, Science & Technology Education*, *9*(4), 371–378.

Pietzner, V. (2014). Computer-based learning in chemistry classes. *Eurasia Journal of Mathematics, Science & Technology Education, 10*(4), 297–311.

Rahmi, U. (2011). *An evaluation of the Indonesian National Examination.* Retrieved November 02, 2014,fromhttps://zh.scribd.com/doc/48858766/An-Evaluation-of-Indonesian-National-Examination-

Reckase, M. D. (2009). *Multidimensional item response theory, Statistics for social and behavioral sciences*. Springer Science &Business Media.doi 10.1007/978-0-387-89976-3 1

Segall, O.D. (1996). Multidimensional adaptive testing, *PSYCHOMETRIKA, 61*(2), 331–354.

Solopos. (2014, April 22).*Jual beli kunci UN sindikat penjual kunci UN beroperasi lintas daerah* [National examination syndicate key answer seller operated nationally]. Retrieved November 11, 2014, from http://www.solopos.com/2014/04/22/jual-beli-kunci-un-sindikat-penjual-kunci-un-beroperasi-lintas-daerah-503932?mobile_switch=mobile

Sulistyo, G.H. (2009). English as a measurement standard in the national examination: Some grassroots voice. *TEFLIN Journal, 20*(1)*,* 1–24.

The Ministry of National Education (2006). *Peraturan menteri pendidikan nasional nomor 23 tanggal 23 mei 2006 about standar kompetensi lulusan (SKL)* [Regulation of national education ministry number 23 date 23 may 2006 about graduate standard competencies]. Jakarta: MONE.

Thomson, N.A., & Weiss, D.J. (2012). A framework for the development of computerized adaptive tests. *Practical Assessment, Research and Evaluation, 16*(1), 1–9.

van der Linden, W.J., & Glas, G.A. (2002). *Computerized adaptive testing: Theory and practice*. New York: Kluwer Academic Publishers.

Wainer, H. (2000). *Computerized adaptive testing*: *A Primer*. 2nd ed., New Jersey: Lawrence Erlbaum Associates.

Wang, C., Chang, H.H., & Boughton, K.A. (2012). Deriving stopping rules for multidimensional computerized adaptive testing. *Applied Psychological Measurement, 37*(2), 99–122. doi: 10.1177/0146621612463422

Wang, H.P., Kuo, B.C., & Chao, R.C. (2011). A multidimensional computerized adaptive testing system for enhancing the Chinese as Second Language proficiency test. *Selected Topics in Education and Educational Technology*, *5,* 245–252

Wang, H.P., Kuo, B.C., Tsai, Y.H., & Liao, C.H. (2012). A CEFR-based computerized adaptive testing system for Chinese proficiency. *The Turkish Online Journal of Educational Technology, 11*(4), 1–12.

Wang, T., Hanson, B. A., & Lau, C. M. (1999). Reducing bias in CAT trait estimation: A comparison of approaches. *Applied Psychological Measurement, 23*, 263–278. doi: 10.1177/01466219922031383

Weiss, D.J. (2011). *Item banking, test development, and test delivery*. Minnesota: University of Minnesota Press.

Wilson, M., & Adams, R. J. (1995). Rasch models for item bundles. *PSYCHOMETRIKA, 60,*181–198.

Wu, M. (2012). *ConQuest Software.* Australia: Australian Software Ltd.

Yao, L. (2012). Comparing the performance of five multidimensional CAT selection procedures with different stopping rules. *Applied Psychological Measurement, 37*(1), 3–23.doi: 10.1177/0146621612455687