REID (Research and Evaluation in Education), 10(1), 2024, 35-49
Available online at: http://journal.uny.ac.id/index.php/reid

Simulation of low-high method in adaptive testing

Rukli[1*]; Noor Azeam Atan[2]
[1]Universitas Muhammadiyah Makassar, Indonesia
[2]Universiti Teknologi Malaysia, Malaysia
[*]Corresponding Author. E-mail: rukli@unismuh.ac.id

ARTICLE INFO

Article History
Submitted:
25 October 2023
Revised:
19 March 2024
Accepted:
16 April 2024

Keywords
low-high method; adaptive testing; item response theory; item difficulty level

Scan Me:

ABSTRACT

The era of disruption significantly engineered a classic testing system into an adaptive testing system where each test taker takes a unique test. However, the carrying capacity of the adaptive testing system engineering is experiencing obstacles in terms of the method of presenting the test questions. The study aims to introduce the low-high adaptive tracking method with the item response theory approach, where the difficulty level of the questions is adapted to the test takers' abilities. The number of test questions in the question bank is 400 questions. Data analysis used the Bilog-MG program. The range of the difficulty level of the questions and the ability level of the test takers was determined [-3.3]. The initialization of the ability of each test taker is set flexibly. The test taker's response uses the pattern of all wrong, all true, and normal responses. The research results show that the low-high method with the IRT approach matches the pattern of the method of giving adaptive test questions. The low-high method requires about 17 test questions to find the ability of the test takers. Another characteristic of the low-high method is that if the responses of the test takers' three to five questions are all correct, then the calculation of divergent abilities is positive, and if the responses of the test takers' three to five questions are all wrong, then the calculation of convergent abilities is negative. Teachers can use the low-high method to design and assemble adaptive tests in schools electronically and manually.

To cite this article (in APA style):
Rukli, R., & Atan, N. (2024). Simulation of low-high method in adaptive testing. *REID (Research and Evaluation in Education), 10*(1), 35-49. doi:https://doi.org/10.21831/reid.v10i1.66922

INTRODUCTION

The shift of the testing system from the classic model to a modern model is due to the needs of the 21st-century climate (Boussakuk et al., 2021) including an increase in the test package, which turned out to be more effective (Lin et al., 2021) as well as online learning (Tripathi et al., 2022). Experts continuously carry out test simulations for adaptive testing system specifications (van der Linden, 2022; Huang et al., 2022). This shows that the adaptive testing system is increasingly studied and elaborated down to the level of test characteristics and test items.

The assembling of the items into tests that have been carried out so far uses a wide range of the difficulty level distribution of the questions or the distribution is heterogeneous (easy, medium, and difficult). There are even more details using more categories: very easy, easy, medium, difficult, and very difficult. More and more categorization is closer to continuous variables, which follows the philosophy of the test that the range of abilities is from infinite negative to infinite positive (Hambleton & Swaminathan, 1985; Hambleton, 1989; Reise et al., 2005).

Conversely, the closer the categorization to categorical or discontinuous variables, the more questions that are not effective in revealing the abilities of test takers with different abilities. The assumption is that the test should have a wide distribution of difficulty levels, but it is not effective so it is better if the test items have a narrow distribution of difficulty levels around the average. This is not following an adaptive testing system where each test taker has a unique ability.

Administering an adaptive testing system uses a procedure that all existing test questions are effective and optimal in measuring the test takers' ability (Kozierkiewicz-Hetmańska & Poniatowski, 2014; Rukli, 2018). If the test taker can do a question correctly, he will be given a question with a higher level of difficulty; if he fails to do it, he will be given a test item with a lower level of difficulty. If test takers with low abilities work on questions with a low difficulty level, then the information about the test takers' abilities is accurate, and if test takers with high abilities work on questions with a high level of difficulty, then the information about the test takers' abilities is also accurate.

Educational and psychological test theories are known as Item Response Theory (IRT) and Classical Test Theory (CTT) (Ma et al., 2020). CTT has the characteristics of test questions depending on the test group so that the statistical values obtained depend on the sample. As a result, if the test group changes, the statistical values on the questions and tests will change. Conversely, in IRT, once the questions have been calibrated, the values obtained will be invariant where they are not affected by the test group so that the item values, known as the item parameters, do not change. The characteristics of the test items refer to the independent parameters of the test group so that they match the adaptive test (Rukli & Hartati, 2011).

The Item Information Function (IIF) is inversely proportional to the square of the standard error of the ability parameter estimate (Hambleton & Swaminathan, 1985; Stenbeck et al., 1992; Hulin et al., 1982). That is, if the IIF is large, the standard error for estimating the ability parameter will be small. If the test taker works on two or more questions where the difference in the standard error of the ability parameter estimate is smaller and the criteria set, then the question is stopped, or there are other criteria so that mathematically, the test taker's ability estimate has been found. This procedure is a stage for making computer-based adaptive tests, or it can also be done manually by assembling questions starting with easier questions.

One concept describes the characteristics of the questions related to the test takers' ability, namely IRT. IRT is in the form of a logistics model. If only paying attention to the difficulty level of the questions related to the test takers' ability, then the one-parameter logistic model is used. When paying attention to the questions' difficulty level and the discrimination of the questions related to the test takers' ability, a two-parameter logistic model is used. When paying attention to the different discrimination powers of the questions, the questions' level of difficulty, and the probability of guessing the questions related to the test takers' ability, a three-parameter logistic model is used. The interpretation of the characteristics of these questions in the form of a multiple-choice test has a different meaning for each model, so it has implications for the interpretation of the test takers' abilities. These three models can be applied to adaptive testing for example (Senge & Hullermeier, 2015; Sineglazov & Kusyk, 2018). The one-parameter logistic model is simple regarding application and concept because it is more suitable for small-scale tests where the number of questions in the question bank is relatively limited but accurate. Besides, the difficulty level of the questions and the level of test takers' ability are 'prioritized' on the same scale.

The engine is the basis for the locomotive direction to move according to the ability of the test takers. The low-high method is the opposite of the high-low method. The low-high method is a triangular branch method with the rule that if the test taker's response is wrong, then an easier question is given; otherwise, if the test taker's response is correct, then a difficult question is given (Hulin et al., 1982). This process is continued until the iteration of the assessment converges where the difference between the two standard errors of the respective difficulty level parameter estimates is smaller or equal to 0.01. The low-high method can be used as a reference in choosing adaptive questions to abilities with several response models to check the test length and the exposure level of the test items. For this reason, the study aims to compare the pattern of adaptive test results with abnormal and normal response patterns. Besides, it aims to find the test length and the questions' exposure level in the normal response model to convergent assessment.

Item Response Theory

Item Response Theory (IRT) places a parameter scale of the level of difficulty of the test items with the parameter level of the ability of the test takers (Halama & Biescad, 2011; DeMars, 2018). The placement of the two gives many further consequences when compared to other parameters, namely discrimination and probability of guessing in a three-parameter logistic model. The - discrimination parameter is the touch point of the tangent line to the logistics model curve. The probability of guessing parameter is the meeting point between the logistic model curve line and the y-axis.

The item difficulty level parameter is a parameter of the item character, which is a feature of all logistical models in IRT. The only one-character model of the item parameter associated with the test taker's ability parameter is the one-parameter logistic model. The one-parameter logistic model equation is presented in Formula (1) as follows.

$$P_i(\theta) = \frac{e^{D(\theta - b_i)}}{1 + e^{D(\theta - b_i)}} \quad \text{..........} \quad (1)$$

Formula (1) shows only two variables that feature, namely, b_i the difficulty level and θ (theta) the ability level of the test taker. The other feature is D = 1.7 as a constant. Using the Excel application, where the two features are added by one constant, makes the logistic model curve one parameter, as shown in Figure 1.

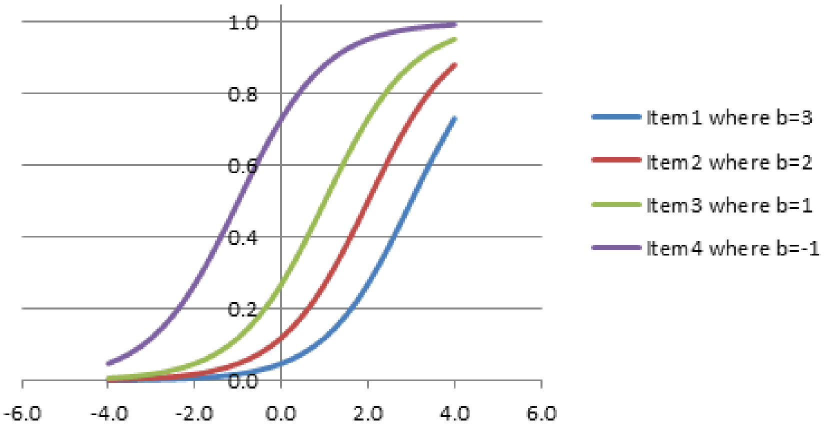

Figure 1. Four Test Items with Different Difficulty Levels

Figure 1 shows four curves of test items with difficulty levels of 3, 2, 1, and -1. The one-parameter logistic model has the same curve; if it is pulled left or right parallel to the x-axis, it will remain the same. In other words, the parameter of the difficulty level of the questions and the parameter of the test takers' ability are on the same scale. The implication is that the two variables of these parameters can be compared directly and functionally. Based on this, the two features can become a theoretical basis for developing an adaptive exam system.

METHOD

Flow Chart

Theoretical studies and the results of previous studies, as well as research objectives, are a reference for designing an adaptive testing system flow chart. The flow chart of the adaptive testing system is shown in Figure 2.

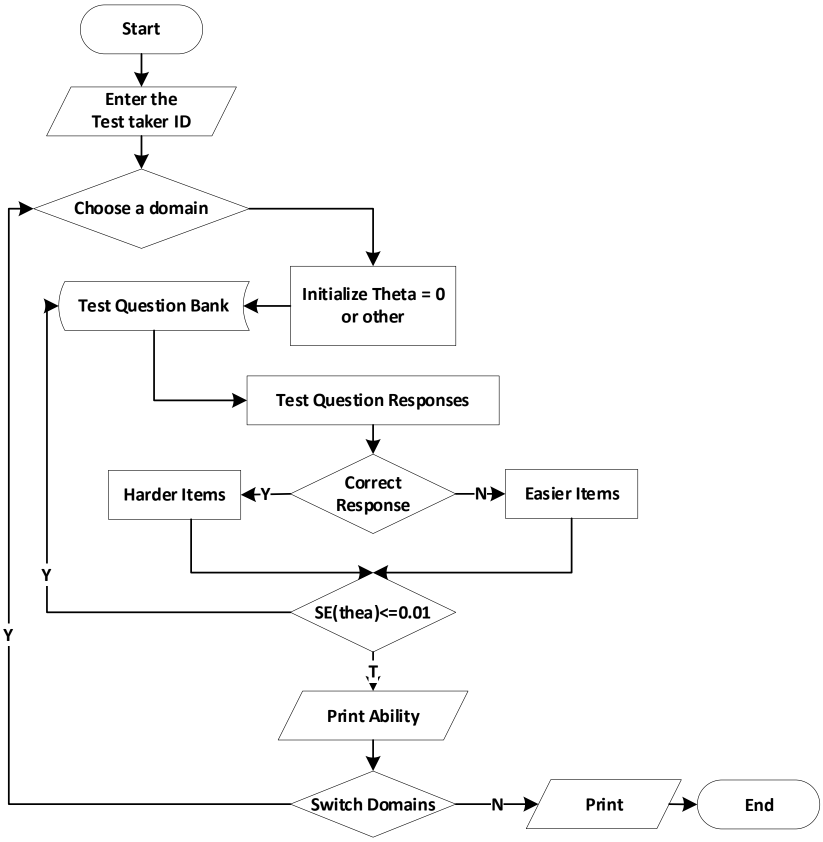

Figure 2. Flow Chart of the Low-High Method

Figure 2 shows the stages of the low-high method of the adaptive testing system. First, the selection of domains can be in the form of subjects to be taken, for example, mathematics, Indonesian, sub-subjects, or themes. However, in this simulation, the domain is subjects, namely mathematics and the Indonesian language. Second, initialization determines the test takers' ability, which is decisive at the next stage of the process. There are various ways to determine the adaptive test stages, for example, choosing a moderate level of difficulty related to the test taker's ability or using extra questions related to the domain to be measured where the results of the ability assessment are the initial initialization of the test taker's ability where each of them has an initial theta itself so that it is quite accurate but less applicable when the ability initialization process requires a lot of questions, especially in the question bank whose number is limited (Triantafillou et al., 2008). In addition, to keep the research data confidential, the participants were anonymized, their participation would not jeopardize their future, and the data obtained were only used for research purposes.

This research uses a theta ability level equal to zero, assuming that this ability is at the midpoint of the ability range from -3 to 3. Likewise, selecting the next item will be closer to the medium difficulty level. This determination is said to match the level of difficulty of the questions with the level of ability (Hulin et al., 1983) and is more or less the same as matching the maximum item information function at the ability level if $c = 0$ and $a = 1$ (one-parameter logistic model). However, this determination is not strictly followed for simulation purposes, but theta may not equal one.

Third, the selection of the next question is carried out using the descending and ascending methods. This method is a decision tree method, which is the simplest and oldest method. These methods vary in terms of both the number of tree branches and the basis for subsequent tracking (Magee, 1964; Sarabia et al., 2021). The minimum number of tree branches is two, and the maximum is not limited, although the determination of the number is following the needs and effectiveness. The need for too much detail and accuracy requires more branching.

The effectiveness of branching is related to tracking the flow, where the more branches there are, the longer the decision-making process, thereby wasting computer/laptop time and space. Besides, the presence of many branches will create many possible holes or dead ends so that instead of being detailed and accurate, it can even happen the other way around. This is the basis for choosing a two-pronged tracking method, namely a low-high method.

Low-high Method

The low-high tracking method is derived from the high-low tracking method. Both methods are two-pronged. The high-low method is a method of tracking up first to higher, then tracking down, and so on until reaching convergent tracking. Low high tracking is a revision of high low tracking even though the changes appear to be just reversed in the direction of the order but have a theoretical test philosophy.

CTT or IRT do not conflict with the philosophy that test takers experience less than normal conditions at the start of the exam anywhere and under any conditions. Even so, these less normal conditions varied among test participants. Such conditions can decrease the test taker's ability. When test takers answer questions, they will also experience interference or fail to answer correctly or perfectly according to their normal abilities. Conversely, test takers who were outliers from the previous review or did not experience any drop or interference at all, then the first question given was lower than their abilities so that the test questions are a prize to being able to answer more questions on the next test. Therefore, the first question when the exam is given is a question that has a lower level of difficulty than the actual ability of each test taker. This is the ideal way. Normally, test takers are given questions at the level of difficulty according to the average ability of the group.

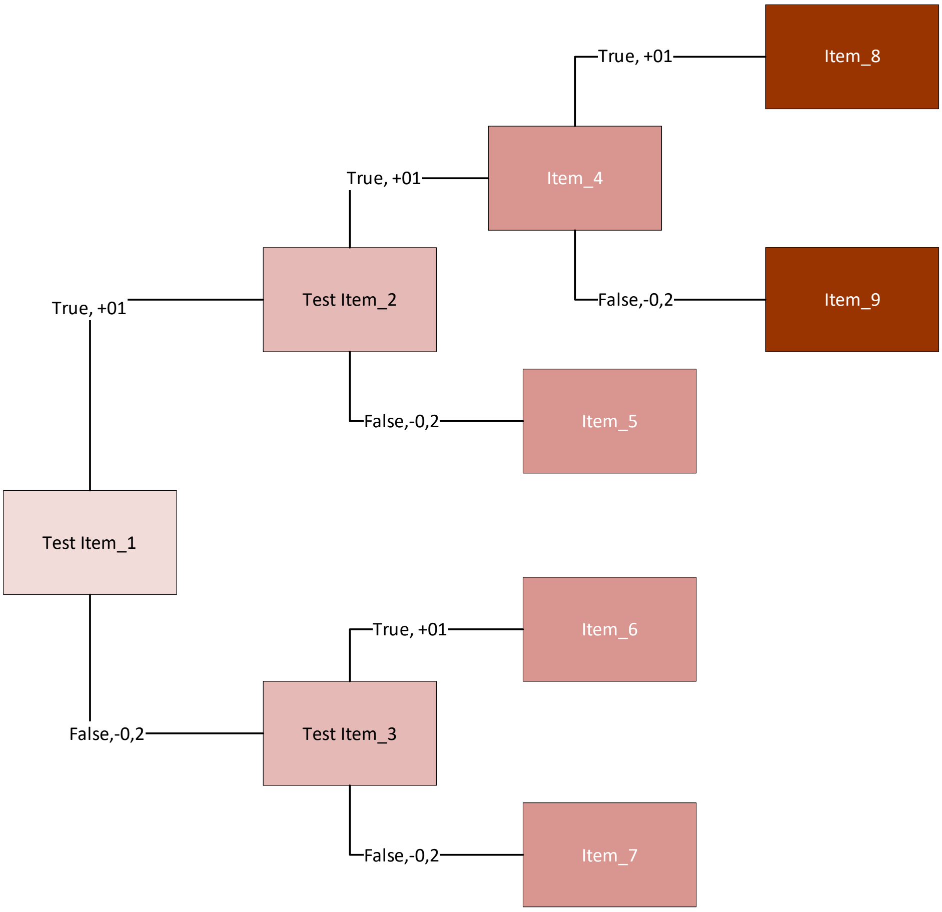

Figure 3. Low-High Method

Besides, the number of branches of the tracking method depends on the next engine, namely how much it goes down if the answer is wrong and how much it goes up if the answer is right. The low-high tracking method has an engine; that is, if a test taker answers incorrectly, the level of difficulty of the lowered test item is greater than equal to 0.2. Conversely, if it is true, then the difficulty level goes up or more than 0.1 (Hulin et al., 1983). This pattern makes the fluctuation tracking method unique and simple. The pattern of the low-high tracking method is shown in Figure 3. The low-high tracking method is the basis for providing adaptive test questions according to the test takers' current abilities.

Fourth, the test termination rule uses three rule conditions. First, test takers respond to question after question until their ability is known. This refers to the standard error of estimating the ability parameter from the results of the responses to two consecutive test questions, which are smaller or equal to 0.01. In other words, if additional questions are given, the test taker's ability does not change significantly so that the system stops and provides a trace report on the test taker's abilities, both quantitatively in the form of numbers and descriptively in the form of graphs. This information is useful for test takers and teachers, namely, test questions that can be done and test questions that students cannot do. This provides further information on remedial materials. Second, test takers have to make the best use of their time. The test taker answers each question within three minutes. The questions related to the domain of mathematics are 40 questions, so the total time allocation is only 120 minutes. If the time limit is exceeded, the exam will be terminated. Third, the test items in the question bank are no longer available for the test takers according to their current abilities.

Fifth, the ability output of the test taker is in that domain. Test takers must choose an available domain and proceed with the process as in the previous steps. Test results for each domain for each test taker can be printed by the test taker at that time, accompanied by questions and traces of his/her ability.

Response Pattern

Response patterns use all true, all false, and normal response patterns. The answer pattern is as follows. First, the pattern of all correct responses means that the test takers correctly answer all the test items given. This pattern reflects the extremely positive response pattern. This informs the test takers that they have abilities above the difficulty level of the questions.

Second, the pattern of all incorrect answers means that the test takers answered incorrectly all of the test items. The wrong pattern all reflects the extremely negative answer pattern. This shows that the test takers' ability is below the difficulty level of the questions.

Third, the normal response pattern means the response pattern, as usual, namely sometimes wrong or sometimes right, with the answer pattern 111¦0110110100¦000 (Linacre & Wright, 1994). However, for simulation purposes, this pattern is not strictly used, so other patterns similar to this pattern remain a simulation reference. The normal pattern and its variations are mostly used to check the test length and the exposure of the test questions from each simulation. All of these answer patterns aim to test the simulation low-high method in giving test questions according to the test takers' ability. These results inform the number of test items (test length) and trace the ability of the test takers and the exposure level of the test items.

Test Length and Item Exposure

Test length and item exposure are features of an adaptive testing system. The test length concerns the number of test questions given to test takers until they converge. The test length is determined by the average number of questions in each of the nine simulations. The number of simulations in this study is 700 times the simulation of the normal response model, while the abnormal response model is only 10 times. Item exposure in a simulation study is the frequency of an item appearing between simulations. This is intended to track the balancing of material for each test taker so that the material is measured optimally.

FINDINGS AND DISCUSSION

Findings

All Correct Response

The pattern of all correct responses indicates that the system is looking for questions with increasing difficulty levels, or there are no items that match, or the assessment stops because the difference in - Standard Error (SE) is less than or equal to 0.01. The last column shows a final score of 145, greater than 100. This value does not naturally occur because the range of abilities of the test takers is defined as [–3.3] or the score is in the range [0. 100]. This can happen if the responses are all correct, and the system will continue to look for higher questions so that no suitable questions are available. Furthermore, if all the responses are correct, the system will continue searching for questions with a higher difficulty level. As a result, no more items may be filled in the question bank. The system has been designed to anticipate these conditions by displaying a warning message. This means that there are no questions that match the test question bank. Another message is that the time for working on the test questions is up, or the assessment has been reached where the test takers' ability is known.

Figure 4 shows the pattern of all correct responses with seven initializations and 10 items, with the score for each simulation being higher. If the number of items increases, the score will be even greater, which is close to infinity. Even so, this is not a problem because, theoretically, the test takers' ability is stretched (-∞.∞). These results indicate a positive divergence. Ten curves out of 10 questions never intersect with other lines. This shows that the simulation results are consistent despite using several theta values. Besides that, the increase in score is in line with the increase in theta value.

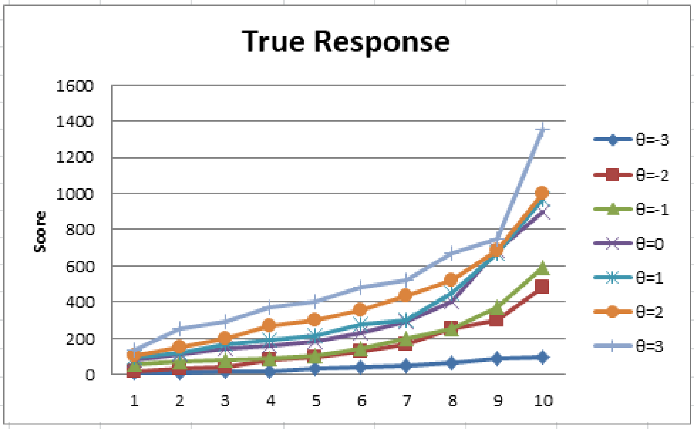

Figure 4. All Correct Response Curve

False Response

Any pattern of wrong answers implies the system is looking for questions of decreasing difficulty. The score column shows the item is smaller than 0, namely a negative score. This can happen if the responses are all wrong, and the system will continue to look for lower questions so that no suitable questions are available. This is not a problem because the ability range of the smallest test takers is not limited even though in simulation, it is limited to -3.

Furthermore, if all the responses are wrong, the system will continue searching for questions with a lower difficulty level. As a result, there may be no more items that fill in the question bank. The system has been designed to anticipate these conditions by displaying a warning message. This means that there are no questions that match the test question bank. Another message is that the time for working on the test questions is up, or the assessment has been reached where the test takers' ability is known.

Figure 5 shows the pattern of all wrong answers with seven initializations: -3, -2, -1-, 0, 1, 2, and 3. There are 10 items used in the simulation. As the number of questions increases, the score decreases to a small to near negative infinity. The simulation shows that the lowest score is -1356 with theta = -3, while the highest score is -103 with theta = 3. If the number of items is added by more than 10, the score will be even greater, close to negative infinity. This result indicates a negative divergence.

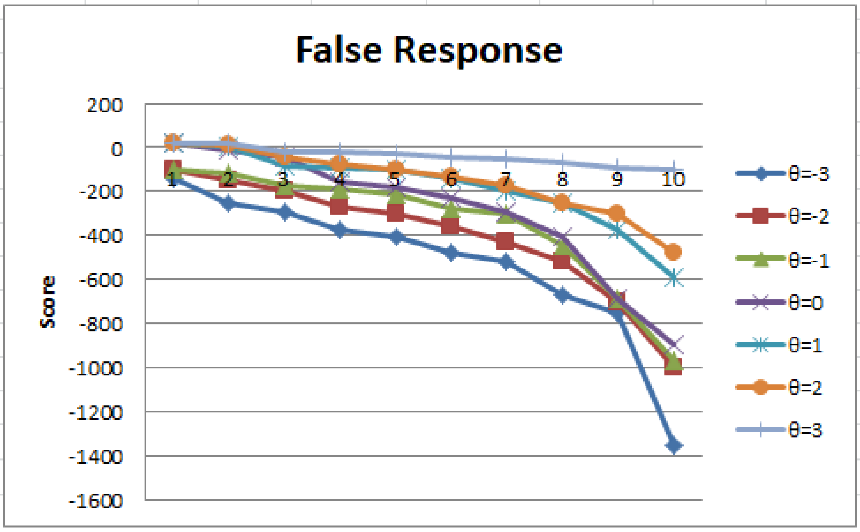

Figure 5. All Wrong Response Curve

Normal Response

The normal response pattern is a response pattern showing that if the difficulty level of the test item is too high compared to the test taker's current ability, the response will be wrong. Conversely, if the difficulty level of the test item is too low compared to the test taker's current ability, the response will be correct. The simulation results show that s answers with varying patterns of normal answers where the test takers were initially wrong, maybe due to their un-familiarity with using computers, anxiety, or low ability compared to the difficulty level of the questions that appear on a computer or laptop screen. However, the second item began to stabilize until the fifth item, for which the response was correct. At that time, the difference in SE had reached a value of 0.008, which was less than or equal to 0.01, so the system stopped.

Table 1 shows the test takers doing a test consisting of seven questions. The final result shows a final score of 84.735 on a [0.100] scale or 2.084 on a [-3.3] scale. Furthermore, if there is a special case where the normal response does not stop, it causes the system to continue looking for questions with a higher difficulty level. As a result, there may be no more items that meet these requirements in the item bank.

Table 1. Patterns for Normal Responses

No.	Id_Item	Theta	b	u	SE	SE Different	Score
1.	104	0	-0.002	0	1.176	0	48.772
2.	120	-0.074	-0.215	1	0.835	0.342	48.477
3.	116	-0.091	-0.113	1	0.681	0.154	51.625
4.	123	0.097	0.065	1	0.599	0.092	60.574
5.	132	0.634	0.166	1	0.535	0.055	75.136
6.	181	1.508	0.322	0	0.513	0.022	79.576
7.	161	1.775	0.043	1	0.505	0.008	84.735

The ability of the test takers (theta) increases if the questions are answered according to the answer key or vice versa. Figure 6 shows a movement of ability from the bottom left to answer the wrong question and then to the right, where the movement takes a non-straight line. The non-straight line indicates the test taker's ability to do the test normally. The more the questions are answered or worked out by the test takers until the assessment converges, the smoother the curve will be traced.

The normal response results show that the final score of seven thetas is no more than 100. Besides that, the score for each item out of 10 increases according to the number of items. In other words, if the previous response is wrong, then to the next question, the system gives a lower item difficulty level so that the score decreases, and vice versa. This shows that the low-high method follows the mechanism of giving adaptive questions according to the current ability of test takers.

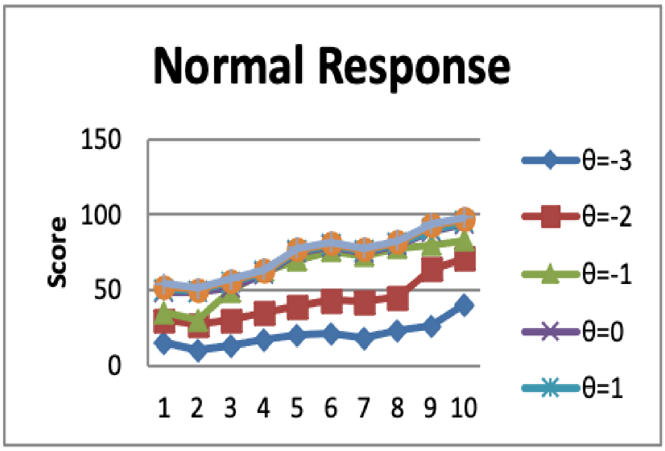

Figure 6. Normal Response from Ten Test Items

The simulation results show that the theta is equal to 0, and other variations of theta values are shown in Table 2. Table 2 shows that the theta values give the same level of difficulty, while when the values vary, the difficulty level values are different. This also applies to the standard error value. The value of the difficulty level and standard error depends on the response pattern. In other words, the response pattern becomes a benchmark for these two values so that the low-high method can be an engine for determining adaptive items to the variance of the test takers' abilities.

Table 3 shows a random sample of nine simulation results from 700 simulations. The average test length during the nine simulations is 17. The test length is less than half of the length of the PTT test, which is 40 questions on the National Examination in Indonesia. The number and types of questions are different, meaning that even though the average number of questions is 17 items, the appearance of the questions is the same. Test questions with ID 221 and ID 225 appear in each of these simulations. The two test questions appeared around the beginning of the exam so that they could open up opportunities for test takers to cheat.

Table 2. Responses to the Nine Test Items

Id Item	Theta	Difficulty Level	Response Pattern	Standard Error
104	0	-0.002	It is all true	1.176
104	0	-0.002	All Wrong	1.176
104	0	-0.002	Normal	1.176
122	-2.134	-2.211	It is all true	0.839
321	-2.238	-2.409	All Wrong	0.197
198	-2.003	-2.151	Normal	1.166
102	1.789	1.681	It is all true	0.102
129	-2.893	-2.906	All Wrong	0.101
271	1.637	0.065	Normal	0.770

Table 3. Question ID and Length in Normal Response

No.	Simulation	Question Id	Question Length
1.	I	123, 126, 221, 225, 004, 005, 129, 132, 133, 284, 254, 255, 331, 170, 192	15
2.	V	010, 111, 221, 225, 219, 108, 177, 105, 091, 227, 288, 192, 120, 129, 203, 222	16
3.	IX	104, 194, 136, 019, 221, 225, 114, 126, 037, 057, 106, 111, 333, 293, 209, 192, 197, 129	18
4.	XX	101, 005, 004, 008, 105, 156, 129, 002, 091, 084, 121, 164, 194, 136, 280, 221, 225	17
5.	XXX	201, 102, 018, 039, 149, 221, 225, 109, 105, 157, 138, 329, 269, 345, 297, 199	16
6.	CCC	104, 120, 116, 123, 132, 372, 238, 324, 221, 225, 213, 305, 232, 222, 197, 322, 352, 189, 191	19
7.	CD	113, 145, 157, 189, 007, 221, 225, 211, 120, 344, 361, 326, 299, 277, 159, 179, 203, 2015	18
8.	DC	210, 221, 225, 347, 321, 003, 016, 082, 017, 186, 100, 102, 203, 291, 304, 356, 364	17
9.	DCC	221, 225, 150, 158, 218, 210, 219, 108, 177, 101, 219, 229, 328, 237, 306, 112, 183	17

Discussion

The adaptive testing system is a spectacular innovation from the classic testing system, namely the paper-and-pencil test (PPT) testing system. PPT conducts item tests strictly where the test makers follow the normal rules, namely the normal curve. The normal curve becomes the PPT reference so that the proportion of questions is based on the level of difficulty of the questions, for example, 30% easy questions, 40% average questions, and 30% difficult questions. This method is theoretically understandable and does not violate existing statistical norms. However, if viewed from the theory of the test, this needs to be removed because it lacks respect for the uniqueness of each test taker.

The uniqueness of each test taker, wherever and whenever, still needs to be considered, including when creating and designing tests, so that there is equality and fairness. One form of inequality and injustice in examinations is passing students who do not pass (Cornelisz et al., 2019). The equivalent concept of the adaptive testing system is to match the level of difficulty of the questions with the level of ability of the test takers on the same scale so that the scores can be compared. Fairness refers to the characteristics of the test takers' ability, not the test items' characteristics.

Referring to the PPT test procedure, namely giving test items to all test takers equally without discriminating. However, in reference to the characteristics of test takers' abilities, the concept of an adaptive testing system, namely giving test items to all test takers according to their abilities, is used. Because each test taker has a unique ability, the adaptive testing system must embody the elements of equality and fairness.

The simulation results of the adaptive testing system use the low-high method for all response simulations following the test theory. That is, the low-high method can be a reference for determining adaptive test items to test takers' abilities with several notes. The first note is the positive divergent assessment. The results of the assessment increased in a dispersive manner when the test takers worked out some of the questions. This means that the more the test items are answered, the more inaccurate the estimate will be. The more inaccuracy shows that the test questions are not functional. If this continues, it can drain the item bank. Positive divergence estimates are not expected to occur in an assessment. Therefore, a positive divergent model estimate is opposed to a convergent estimate.

The second note is the negative divergent assessment. The assessment results fell out of focus when the test takers with wrong responses answered more than two questions. The more the test items are done, the more the estimates will spread downward so that they are away from the target point. If this continues, it can drain the item bank and positive divergence. Negative divergence estimates are not expected to occur in an assessment. Therefore, the negative divergent model estimates are contrary to the focus of the convergent assessment results.

The third note is a convergent approximation. The assessment results focus on when the test takers answer some of the questions. This means that the more the test items are answered, the more accurate the estimate will be. The higher accuracy shows that the functional test questions are focused more on the estimated target. If this is continued, it can save the question bank item bank. For this reason, convergent estimation is expected to occur in an estimation. Therefore, the assessment of the convergent model follows the rules of test theory so that the convergent assessment is used to estimate the test taker's ability level.

Convergent assessment corresponds to the concepts of test reliability and test validity (Saidi & Siew, 2019; Balachandran et al., 2021). Reliability refers to the consistency of an item to construct a test kit. Test reliability is related to the information function of items and tests in IRT (Alnasraween et al., 2021). The test information function is the result of the accumulation of the test item information function (Hoshino, 2001; Boone & Staver, 2020; Frey, 2018). The information function can be applied to track the high-level anxiety of medical students (Zhang et al., 2021).

The low-high method provides adaptive test questions to test takers following the concept of test theory. These results are in line with the decision tree of Rodríguez-Cuadrado et al. (2020), Bayesian method (van der Linden, 1998; Veldkamp & Matteucci, 2013), Kullback-Leibler information with a posterior distribution (Victor et al., 2020), and Sympson-Hetter method (Han, 2018). In addition, the adaptive testing system can add advanced functions of polytomous item response models, weighted likelihood estimation methods, and content balancing methods (Seo & Choi, 2020).

Moreover, adaptive testing systems can be coupled with formative tests, which serve as efficient educational evaluation tools for personalized distance learning services (Choi & McClenen, 2020). Likewise, adaptive testing systems greatly reduce the number of test items and time without losing measurement accuracy (Xu et al., 2020) and allow individual assessment for longitudinal monitoring (Lai et al., 2017).

The number of test items in convergence assessment is smaller, namely around 12, when compared to that in PPT, which is around 40. This is in line with previous studies of more than half (Li et al., 2020) as well as a more active and efficient adaptive testing system (Han, 2018). Besides, the adaptive testing system is a method of selecting items and transition criteria so that it can increase the accuracy of estimating certain latent variables to different levels (Bao et al., 2021). Therefore, the adaptive testing system with the low-high method becomes a reference for teachers in designing adaptive tests with questions from PPT. This is not an obstacle because, in essence, PPT testing can be made adaptive to support teaching in the classroom (Chang, 2015). However, the low-high method cannot control the exposure of the test items, especially at the beginning of the test.

Item test exposure is the frequency with which an item appears in an exam or between exams from several test takers. If this occurs several times, it will reduce the confidentiality of the test items or can cause participants to cooperate. Factors influencing item exposure are the psychometric properties of the items in the item pool (Revuelta & Ponsoda, 1998) and the composition of the items in the range of the desired ability scale (Ozturk & Dogan, 2015). There are many studies on item exposure control, for example, the continuous a-stratification index by Huebner et al. (2018), Simpson and Hetter by van der Linden and Veldkamp (2004) and Chen et al. (2008), and the knowledge-based approach (Doong, 2009). It needs to be studied more in detail because it can weaken the balance of items appearing in adaptive tests. Thus, further research can use other methods, for example, maximum likelihood estimation, Bayesian, or heuristic either separately or in combination with low-high methods to control the exposure of the test item.

CONCLUSION

The low-high method requires around 17 questions until the test takers' ability estimates converge. If the test takers' responses in three to five questions are all correct, then the calculation of divergent ability is positive. Conversely, if the test takers' responses successively three to five questions are all wrong, then the calculation of divergent ability is negative. Teachers can construct test items into tests by using the low-high rule to more accurately measure test takers' ability even though the number of questions is small. The low-high method cannot control the exposure level of the test items, so further research can use other methods, for example, maximum likelihood estimation, Bayesian, or heuristic, either separately or in combination with low-high methods. In addition, the number of items in the item bank is very small, namely, less than 1000. The number of such items is questionable regarding the representation of the material being represented. Therefore, future studies can make tests with narrower domain groupings, not subject groupings but the sub-subject-matter. Likewise, care should be taken to widen the range of item difficulty levels, namely wider than the range [-3.3], so that divergent values can be controlled. Thus, the teacher can use the low-high method following the research results. However, if the questions given are not varied according to the variations in the abilities of the test takers, the low-high method can be combined with other methods so that negative and positive divergent estimation results can be avoided.

ACKNOWLEDGMENT

The authors would like to thank the Chancellor of Muhammadiyah Makassar University who has provided internal research grant assistance so that this research can achieve the expected results.

DISCLOSURE STATEMENT

The authors hereby declare that they have no potential conflicts of interest in writing this article.

REFERENCES

Alnasraween, M. S., Al-Mughrabi, A. M., Ammari, R. M., & Alkaramneh, M. S. (2021). Validity and reliability of eight-grade digital culture test in light of item response theory. *Cypriot Journal of Educational Sciences, 16*(4), 1816-1835. https://doi.org/10.18844/cjes.v16i4.6034

Balachandran, A. T., Vigotsky, A. D., Quiles, N., Mokkink, L. B., Belio, M. A., & Glenn, J. M. K. (2021). Validity, reliability, and measurement error of a sit-to-stand power test in older adults: A pre-registered study. *Experimental Gerontology, 145,* 111202. https://doi.org/10.1016/j.exger.2020.111202

Bao, Y., Shen, Y., Wang, S., & Bradshaw, L. (2021). Flexible computerized adaptive tests to detect misconceptions and estimate ability simultaneously. *Applied Psychological Measurement, 45*(1), 3-21. https://doi.org/10.1177/0146621620965730

Boone, W. J., & Staver, J. R. (2020). Test information function (TIF). In *Advances in Rasch analyses in the human sciences* (pp. 39–55). Springer International Publishing. https://doi.org/10.1007/978-3-030-43420-5_4

Boussakuk, M., Bouchboua, A., El Ghazi, M., El Bekkali, M., & Fattah, M. (2021). Design of computerized adaptive testing module into our dynamic adaptive hypermedia system. *International Journal of Emerging Technologies in Learning*, *16*(18), 113–128. https://doi.org/10.3991/ijet.v16i18.23841

Chang, H. H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika, 80,* 1–20. https://doi.org/10.1007/s11336-014-9401-5

Chen, S. Y., Lei, P. W., & Liao, W. H. (2008). Controlling item exposure and test overlap on the fly in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology, 61*(2), 471-492. https://doi.org/10.1348/000711007X227067

Choi, Y., & McClenen, C. (2020). Development of adaptive formative assessment system using computerized adaptive testing and dynamic bayesian networks. *Applied Sciences (Switzerland), 10*(22), 8196. https://doi.org/10.3390/app10228196

Cornelisz, I., Meeter, M., & van Klaveren, C. (2019). Educational equity and teacher discretion effects in high stake exams. *Economics of Education Review, 73,* 101908. https://doi.org/10.1016/j.econedurev.2019.07.002

DeMars, C. E. (2018). Classical test theory and item response theory. In P. Irwing, T. Booth, & D. J. Hughes (eds.), *The Wiley handbook of psychometric testing* (pp. 49–73). Wiley. https://doi.org/10.1002/9781118489772.ch2

Doong, S. H. (2009). A knowledge-based approach for item exposure control in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 34*(4), 530-558. https://doi.org/10.3102/1076998609336667

Frey, B. B. (2018). Test information function. In *The SAGE encyclopedia of educational research, measurement, and evaluation.* SAGE Publications, Inc. https://doi.org/10.4135/9781506326139.n694

Halama, P., & Biescad, M. (2011). Measurement of psychotherapy change: Comparison of classical test score and IRT based score. *Ceskoslovenska Psychologie: Casopis Pro Psychologickou Teorii a Praxi, 55*(5), 400-411.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147–200). Macmillan Publishing Co, Inc; American Council on Education.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Springer Dordrecht. https://doi.org/10.1007/978-94-017-1988-9

Han, K. C. T. (2018). Components of the item selection algorithm in computerized adaptive testing. *Journal of Educational Evaluation for Health Professions, 15,* 7. https://doi.org/10.3352/jeehp.2018.15.7

Hoshino, T. (2001). A test information function for linear combinations of traits without nuisance traits in multidimensional item response theory. *Japanese Journal of Educational Psychology, 49*(4), 491-499. https://doi.org/10.5926/jjep1953.49.4_491

Huang, H. T. D., Hung, S. T. A., Chao, H. Y., Chen, J. H., Lin, T. P., & Shih, C. L. (2022). Developing and validating a computerized adaptive testing system for measuring the English proficiency of Taiwanese EFL university students. *Language Assessment Quarterly, 19*(2), 162-188. https://doi.org/10.1080/15434303.2021.1984490

Huebner, A., Wang, C., Daly, B., & Pinkelman, C. (2018). A continuous a-stratification index for item exposure control in computerized adaptive testing. *Applied Psychological Measurement, 42*(7), 523-537. https://doi.org/10.1177/0146621618758289

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement.* Dow Jones-Irwin.

Hulin, C. L., Drasgow, F., & Komocar, J. (1982). Applications of item response theory to analysis of attitude scale translations. *Journal of Applied Psychology*, *67*(6), 818–825. https://doi.org/10.1037/0021-9010.67.6.818

Kozierkiewicz-Hetmańska, A., & Poniatowski, R. (2014). An item bank calibration method for a computer adaptive test. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8397 LNAI* (Part 1, pp. 375–383). https://doi.org/10.1007/978-3-319-05476-6_38

Lai, J. S., Beaumont, J. L., Nowinski, C. J., Cella, D., Hartsell, W. F., Han-Chih Chang, J., Manley, P. E., & Goldman, S. (2017). Computerized adaptive testing in pediatric brain tumor clinics. *Journal of Pain and Symptom Management,* *54*(3), 289-297. https://doi.org/10.1016/j.jpainsymman.2017.05.008

Li, Z., Cai, Y., & Tu, D. (2020). A new approach to assessing shyness of college students using computerized adaptive testing: CAT-shyness. *Journal of Pacific Rim Psychology, 14.* https://doi.org/10.1017/prp.2020.15

Lin, Z., Chen, P., & Xin, T. (2021). The block item pocket method for reviewable multidimensional computerized adaptive testing. *Applied Psychological Measurement*, *45*(1), 22-36. https://doi.org/10.1177/0146621620947177

Linacre, J. M., & Wright, B. D. (1994). Dichotomous infit and outfit mean-square fit statistics. *Rasch Measurement Transactions, 8*(2). https://www.rasch.org/rmt/rmt82a.htm

Ma, W., Minchen, N., & de la Torre, J. (2020). Choosing between CDM and unidimensional IRT: The proportional reasoning test case. *Measurement: Interdisciplinary Research and Perspectives, 18*(2), 87-96. https://doi.org/10.1080/15366367.2019.1697122

Magee, J. F. (1964). Decision trees for decision-making. *Harvard Business Review.* https://hbr.org/1964/07/decision-trees-for-decision-making

Ozturk, N. B., & Dogan, N. (2015). Investigating item exposure control methods in computerized adaptive testing. *Kuram ve Uygulamada Egitim Bilimleri, 15*(1), 85-98. https://doi.org/10.12738/estp.2015.1.2593

Reise, S. P., Ainsworth, A. T., & Haviland, M. G. (2005). Item response theory: Fundamentals, applications, and promise in psychological research. *Current Directions in Psychological Science*, *14*(2), 95–101. https://doi.org/10.1111/j.0963-7214.2005.00342.x

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35*(4), 311-327. https://doi.org/10.1111/j.1745-3984.1998.tb00541.x

Rodríguez-Cuadrado, J., Delgado-Gómez, D., Laria, J. C., & Rodríguez-Cuadrado, S. (2020). Merged Tree-CAT: A fast method for building precise computerized adaptive tests based on decision trees. *Expert Systems with Applications, 143,* 113066. https://doi.org/10.1016/j.eswa.2019.113066

Rukli, R. (2018). Analysis of mathematical problem with group wise for computerized adaptive testing. *Daya Matematis: Jurnal Inovasi Pendidikan Matematika*, *6*(1), 96-104. https://ojs.unm.ac.id/JDM/article/view/5600

Rukli, R., & Hartati. (2011). Penerapan sistim pendukung keputusan dalam sistem pengujian computerized adaptive testing. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *5*(3), 71–81. https://doi.org/10.22146/ijccs.5215

Saidi, S. S., & Siew, N. M. (2019). Reliability and validity analysis of statistical reasoning test survey instrument using the Rasch measurement model. *International Electronic Journal of Mathematics Education*, *14*(3), 535-546. https://doi.org/10.29333/iejme/5755

Sarabia, J. M., Roldan, A., Henríquez, M., & Reina, R. (2021). Using decision trees to support classifiers' decision-making about activity limitation of cerebral palsy footballers. *International Journal of Environmental Research and Public Health, 18*(8), 4320. https://doi.org/10.3390/ijerph18084320

Senge, R., & Hullermeier, E. (2015). Fast Fuzzy Pattern tree learning for classification. *IEEE Transactions on Fuzzy Systems*, *23*(6), 2024–2033. https://doi.org/10.1109/TFUZZ.2015.2396078

Seo, D. G., & Choi, J. (2020). Introduction to the LIVECAT web-based computerized adaptive testing platform. *Journal of Educational Evaluation for Health Professions, 17.* https://doi.org/10.3352/JEEHP.2020.17.27

Sineglazov, V. M., & Kusyk, A. V. (2018). Adaptive testing system based on the Fuzzy logic. *Electronics and Control Systems*, *2*(56), 85-91. https://doi.org/10.18372/1990-5548.56.12941

Stenbeck, M., Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1992). *Fundamentals of item response theory*. SAGE Publications.

Triantafillou, E., Georgiadou, E., & Economides, A. A. (2008). CAT-MD: Computerized adaptive testing on mobile devices. *International Journal of Web-Based Learning and Teaching Technologies (IJWLTT), 3*(1), 13-20. https://doi.org/10.4018/jwltt.2008010102

Tripathi, A. M., Kasana, R., Bhandari, R., & Vashishtha, N. (2022). Online examination system. In Y. D. Zhang, T. Senjyu, C. So-In, & A. Joshi (eds.) *Smart trends in computing and communications. Lecture notes in networks and systems* (vol. 286). Springer Singapore. https://doi.org/10.1007/978-981-16-4016-2_67

van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, *63*(2), 201–216. https://doi.org/10.1007/BF02294775

van der Linden, W. J. (2022). Review of the shadow-test approach to adaptive testing. *Behaviormetrika*, *49*, 169–190. https://doi.org/10.1007/s41237-021-00150-y

van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics*, *29*(3), 273–291. https://doi.org/10.3102/10769986029003273

Veldkamp, B. P., & Matteucci, M. (2013). Bayesian computerized adaptive testing. *Ensaio: aval. pol. públ. Educ., 21*(78). https://doi.org/10.1590/S0104-40362013005000001

Victor, V. M., Farias, J. S., Freire, V., Ruela, A. S., & Delgado, K. V. (2020). ALICAT: A customized approach to the item selection process in computerized adaptive testing. *Journal of the Brazilian Computer Society, 26,* 4. https://doi.org/10.1186/s13173-020-00098-z

Xu, L., Jin, R., Huang, F., Zhou, Y., Li, Z., & Zhang, M. (2020). Development of computerized adaptive testing for emotion regulation. *Frontiers in Psychology, 11,* 561358. https://doi.org/10.3389/fpsyg.2020.561358

Zhang, C., Wang, T., Zeng, P., Zhao, M., Zhang, G., Zhai, S., Meng, L., Wang, Y., & Liu, D. (2021). Reliability, validity, and measurement invariance of the general anxiety disorder scale among Chinese medical university students. *Frontiers in Psychiatry, 12,* 648755. https://doi.org/10.3389/fpsyt.2021.648755