International Journal of Pedagogy and Teacher Education

**Journal Homepage**: jurnal.uns.ac.id/ijpte

# Assessing Language Reasoning Skills of Indonesian Students Using Computerized Adaptive Testing

**Memet Sudaryanto[1*], Muhammad Nur Yasir Utomo[2]**

[1]Department of Indonesian Language and Literature Education, Faculty of Humanities, Universitas Jenderal Soedirman, [2]Department of Computer and Informatics, Politeknik Negeri Ujung Pandang, Indonesia

**ARTICLE INFO**

***Article History***
Received : June 30, 2024
1st Revision : July 11, 2024
Accepted : July 20, 2024
Available Online : July 25, 2024

***Keywords:***
reasoning;
computerized adaptive test;
language;
assessment;

***Corresponding Author***
Email address:
memet.sudaryanto@unsoed.ac.id

**ABSTRACT**

Language, as the fundamental means of communication, represents the symbolization of thoughts conveyed to others. Human understanding of the structure of messages relies on the speaker's language reasoning ability. This ability can be measured from the simplest to the most complex stimuli. Traditionally, assessing language reasoning has been done through written tests, which require extensive preparation and are time-consuming. This study proposes a model for measuring language reasoning ability using a Computerized Adaptive Test (CAT). The CAT adjusts the difficulty of questions in real time based on the participant's responses. If a participant answers correctly, the system presents a more challenging question. Conversely, the system selects an easier question if the participant answers incorrectly. This adaptive approach ensures a tailored and efficient assessment experience, accurately measuring the participant's abilities. The research began by developing a valid and reliable language reasoning test instrument and its quadrant class. This included determining the starting, jumping, and stopping points, culminating in the CAT design. The results of the CAT proposed in this study can map basic language reasoning skills, starting from understanding the concept of facts, applying linguistic rules according to the agreement formed in the read clauses, breaking down information into more specific forms, judging the values of ideas, combining word selection, ideas formation, and context, analogy thinking, and comparative thinking. The analysis revealed that the participants' dominant ability was in comparative thinking, which involves comparing language forms, conditions, settings, and messages in written discourse. Moreover, the CAT system proposed in this study was proven to speed up the testing process while enabling students to complete the tests according to their abilities.

**How to cite:** Sudaryanto, M., & Utomo, M. N. Y. (2024). Assessing language reasoning skills of Indonesian students using computerized adaptive testing. *International Journal of Pedagogy and Teacher Education, 8*(1),12–25. https://doi.org/10.20961/ijpte.v8i1.89913

## 1. INTRODUCTION

As a fundamental means of communication, language represents a symbol of thought conveyed through individual messages. Effective communication depends on accurately delivering and comprehending these messages, evident in both spoken and written forms. For instance, a student explaining a scientific concept clearly or a professional writing a detailed report demonstrates how language skills reflect logical thinking and reasoning abilities (Sudaryanto et al., 2020).

Various factors, including social, psychological, and neurobiological aspects, influence an individual's communication needs and ability to acquire, use, and understand language. Frequent communication enhances language skills, from vocabulary mastery to accurately interpreting messages. Reasoning is key to understanding these messages, enabling individuals to optimize unclear information. For example, when reading a complex text, reasoning helps logically reconstruct missing information (Leighton, 2006).

Logical reasoning skills are critical in academic settings, where clear and precise language use is essential. Without strong reasoning abilities, writing scientific papers can lead to bias, errors, and ambiguities (Sudaryanto et al., 2019a). This study addresses students' challenges in articulating complex ideas clearly and concisely in scientific writing. Solving this problem is crucial because effective communication of complex ideas is essential for academic and professional success, ensuring that research findings are accurately conveyed and understood.



p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://dx.doi.org/10.20961/ijpte.v8i1.89913

This study aims to evaluate students' linguistic reasoning abilities using a CAT, focusing on Indonesian students from various academic disciplines, including science, technology, and social humanities. Effective language reasoning involves structuring ideas logically and presenting them in an organized, straightforward manner. This study assessed students' ability to understand language analogies and equivalent meanings and interpret discourses. These skills are crucial for mapping creativity in response to stimuli (Triyono et al., 2017). Scientific competence in language reasoning involves processing information and applying linguistic functions in social contexts (Kavanagh & Szweda, 2017). Logical thinking is essential for analyzing situations and developing reasonable solutions. For example, a logical thinker in a debate would observe and analyze facts before concluding, ensuring their arguments are well-founded.

Assessing reasoning skills in language is vital to understanding students' development in using language as a cognitive tool (Dove, 2014). The learning process and outcome assessments should be aligned to measure competencies accurately. Logical thinking, characterized by responding appropriately to linguistic stimuli, is a hallmark of formal reasoning skills (Clark, 2007; Greenspan & Shanker, 2009). Unfortunately, many people overlook the importance of language for sharpening logical thinking. Knowing how to quickly find ideas, respond appropriately, and present accurate facts is critical. CAT is an effective assessment method that adjusts question difficulty based on the participant's responses, making it more accurate and efficient in measuring logical thinking skills (Larson & Madsen, 2013).

In a CAT, questions are presented individually, and the difficulty level adjusts according to the participant's answers. For example, if a student answers a question correctly, the next question will be more challenging; if answered incorrectly, the following question will be easier. This process continues until a test score accurately reflects the participant's ability is obtained (Ramadhan et al., 2020). Developing a CAT requires evaluating components such as the Item Response Theory (IRT) model, question bank, initial item selection, ability estimation method, item selection procedure, and termination rules (Green et al., 1984; Kingsbury & Zara, 1989). Quality assessment instruments must be empirically validated using IRT, which considers difficulty, discrimination, and guessing parameters (Hambleton & Swaminathan, 1991; Retnawati, 2015). A high-quality language test instrument meets relevance, consistency, practicality, and effectiveness criteria. Content validity, ensured through expert judgment, is crucial for creating instruments that accurately measure the intended competencies. Teachers often need more emphasis on these quality criteria when preparing assessment instruments, highlighting the need for proper guidelines and references (Retnawati, 2015).

Based on previous studies highlighting the importance of valid and reliable assessment instruments, this paper aims to contribute by emphasizing the critical role of logical thinking and reasoning in language use. Specifically, this paper seeks to enhance understanding of how these cognitive abilities can be effectively measured by developing and implementing a CAT.

## 2. MATERIAL AND METHOD

This study aims to develop a CAT to measure the linguistic reasoning ability of students in Indonesia. Using a quantitative approach, the study assesses language reasoning skills, starting from sample adequacy testing, validity testing, and reliability estimation (Weiss & Kingsbury, 1984). The sampling technique was stratified random sampling from student populations in eastern, central, and western Indonesia (Lane et al., 2015). The research approach also included a development method, detailed step-by-step for computerized adaptive testing techniques. The research comprised three main steps: needs analysis, data preparation, and the design of computerized adaptive testing techniques. Each step is described as follows:

### *Technical Specification*

The technical specification begins with a needs analysis to identify the hardware and software requirements for the system. To implement a computerized adaptive test, the necessary hardware includes a server with a minimum of a 2 Core CPU, 4 GB of RAM, and 60 GB of SSD storage. This hardware configuration is essential for running the required software components: CodeIgniter version 4 as the framework for the architecture code system, Bootstrap version 4 as the framework for the user interface system, Apache Server as the web server system to ensure the web system functions, and MySQL Database Server for user data storage, a question bank, and other necessary data. All software components are deployed on the hardware server, which runs on the Ubuntu version 20.04 Linux operating system. This setup ensures optimal performance and reliability for the system's operations.


p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://dx.doi.org/10.20961/ijpte.v8i1.89913

***Data Preparation***

The study involved testing the instrument questions, which consisted of nine dimensions. The questions were divided into five categories or quadrants: very easy, easy, moderate, difficult, and very difficult. Each quadrant contained 1000 questions. These questions were based on a linguistic intelligence test specified in the language reasoning indicators. After reconstructing the items into these dimensions, they formed the "mother test of linguistic-logical thinking skills." Each indicator underwent a sampling test to determine whether the items developed measured the indicators based on the designed constructs according to the assessment's purpose (Kingsbury & Zara, 1989; Wainer et al., 2000).

The test instrument's validity was assessed through content validation, following Aiken's method (Aiken, 1985). Construct validity was confirmed through confirmatory factor analysis. Reliability estimation was conducted using Cronbach's alpha and the item information function with the Item Response Theory (IRT) approach. Each item was grouped based on the test construction and indicators and then tested for unidimensionality to meet the IRT prerequisites. Subsequently, a local independence test ensured that the developed test items were not interrelated and that measurements between test items were independent. Invariance testing confirmed that item characteristics remained consistent despite variations in the examinee subpopulations. After validating the IRT assumptions, the ability measurement process determined the test items' difficulty levels.

***System Design***

The focus of further research was on developing the computerized adaptive test system. This process included several designs: use-case design, algorithm design, adaptive test processes, and database design. These designs served as guidelines for the development of the adaptive test system. Details of each design are explained as follows:

***Use-case Design***

The use-case design defined two user roles: admin and participant. The admin accessed user management, participants, questions, answers, and reporting results. Participants were responsible for registering, taking the test, and viewing their results. **Figure 1** below shows the proposed use-case system design:

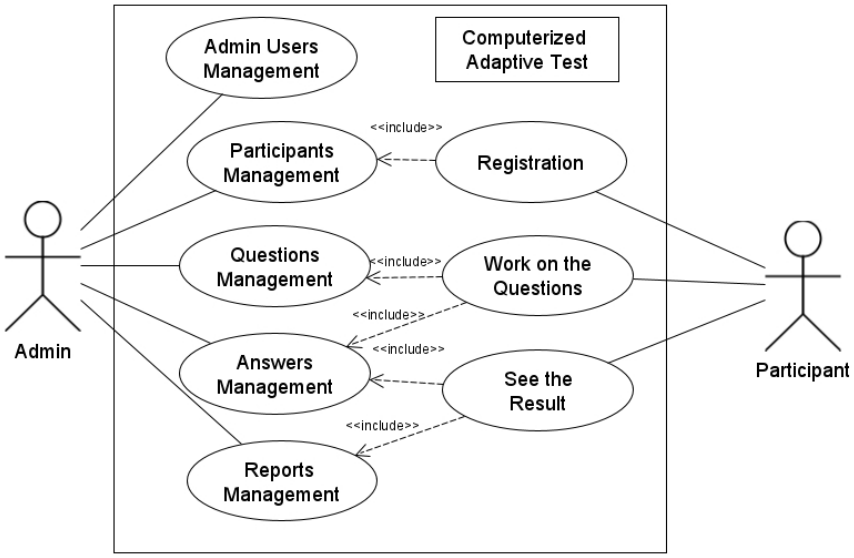


**Figure 1.** Usecase Diagram of Computerized Adaptive Test System

***Design of Adaptive Test Algorithm***

The flowchart design outlines the system's operation for examinees. The process starts with participant registration, logging in, beginning the test, and viewing the results. **Figure 2** below illustrates the running flowchart of the computerized adaptive test system:


p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://dx.doi.org/10.20961/ijpte.v8i1.89913

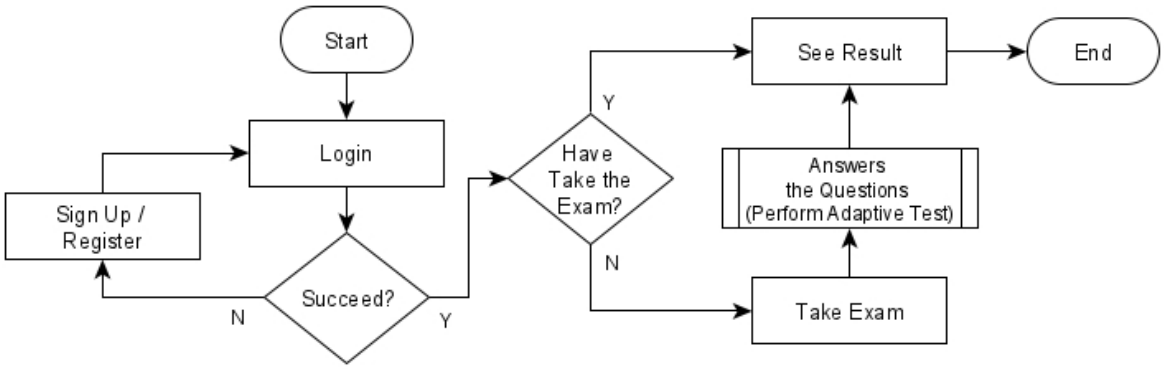


**Figure 2.** Flowchart of the Computerized Adaptive Test System

Figure 2 illustrates the system's operation from registration to logging in and starting the exam. The question difficulty level adapts to the examinee's ability during the exam. The following algorithm adjusts the difficulty level:

**Algorithm 1:** Adaptive Test

**Input** : $\{q_n, d_n\}_{n=1}^{l}$ ; ▷ 5000 Questions, Difficulty Degree
**Output:** Scores

```
Score = 0
curWrong = 0
wrongLimit = 5
for curWrong to wrongLimit do
    User Input : {a=answer, idq=question id, cdd=current difficulty degree}
    if a, idq and cdd is empty then
        // first question
        cdd = 0.000
        ad = Random(allQuestion with d=0.000)
    else
        // not first question
        if a for idq is correct then
            cdd = cdd + 0.005
            ad = Random(allQuestion with d=cdd)
        else
            curWrong +=1
            cdd = cdd - 0.002
            ad = Random(allQuestion with d=cdd)
        end
    end
    Score = ((cdd-(-5)/10))*100
    show question: ad
    show answer option
    if answer choosen then
        continued;
    end
end
```

**Figure 3.** Adaptive Test System Algorithm

Figure 3 illustrates the adaptive test algorithm proposed in this study. The process begins by initializing the input with 5000 questions and setting the output as the CDD (cumulative difficulty distribution) difficulty level. Examinees start with questions at a neutral difficulty level of 0.000. For each correctly answered question, the difficulty level score increases by 0.005. Conversely, for each incorrect answer, the difficulty score decreases by 0.002.


p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://dx.doi.org/10.20961/ijpte.v8i1.89913

As the difficulty level changes, the algorithm selects questions randomly at the corresponding difficulty level. This process operates recursively with a saturation/error limit of five. If the examinee makes five mistakes, the exam process concludes, and the score reflects the last successfully achieved difficulty level by the participant.

***Database Design***

The database design is grounded in the existing flowchart design and adaptive test algorithms. The design uses the Entity-Relationship Diagram (ERD) format. Utilizing the ERD database system, assessments through CAT can be more easily designed and implemented because the ERD visually represents the relationships between data objects and their correlations. Using notations, symbols, and charts, ERDs help organize data structures and relationships (**Figure 4**). This method is highly representative of relational database design.

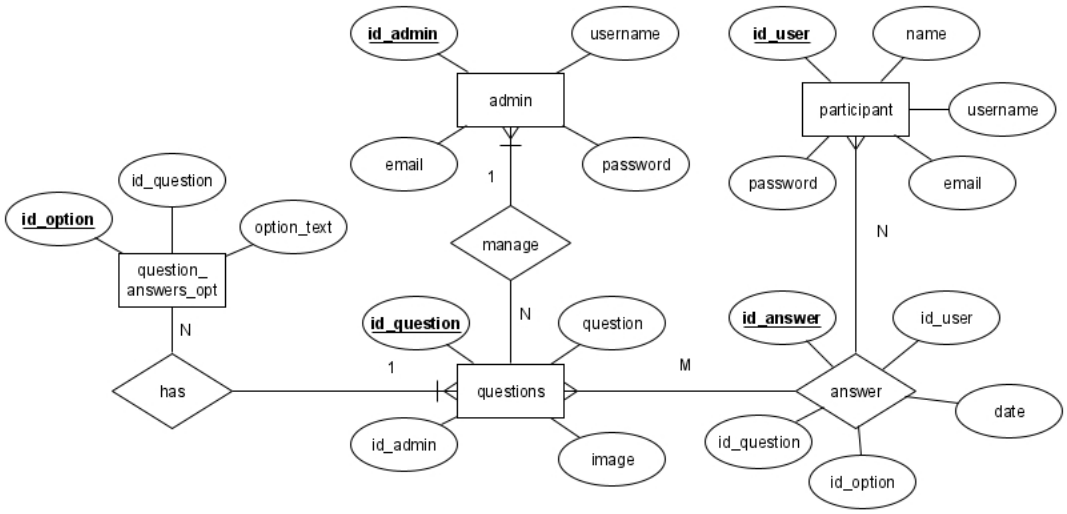


**Figure 4.** ERD Design of System

## 3. FINDINGS AND DISCUSSION

***Development of Test Instruments***

Constructing reasoning abilities involves forming new mental representations through transforming information by a complex interaction of mental attributes, including consideration, abstraction, reasoning, depiction, logical problem-solving, concept formation, creativity, and intelligence. Thinking is an activity that involves gathering ideas or information by connecting pieces of information with the problem at hand.

The test instrument was adapted from general intelligence test questions and improved according to the developed test indicators. Validation based on Aiken's value demonstrated that the reasoning skills of all items were valid, with the smallest Aiken value being 0.85 and the largest at 1.0. Content validity proof with Aiken established a validity score limit of 0.72 for each item based on 9 experts and 5 assessment criteria (Aiken, 1985). A KMO value close to 1 indicates sufficient sample adequacy for further testing, with a KMO value over 0.5 being acceptable. Subsequent confirmatory analysis was performed using Lisrel for Windows (Student) 9.2, as shown in **Table 1**.

The confirmatory analysis of the language reasoning test instrument effectively measured the logical thinking ability of both scientific and social science students (Brown, 2015; Brown & Moore, 2012). Following the measurement model formation using CFA, the relationship between latent variables and indicator variables was analyzed, including testing whether each manifest variable/indicator accurately measured the latent variable. The adequacy of the validity value indicated that the observed variables effectively measured the latent variables with high accuracy. Cronbach's alpha was used to estimate the reliability of the test instrument. An alpha value greater than 0.7 indicates sufficient reliability, while an alpha value greater than 0.87 suggests that all items are reliable and that the tests consistently exhibit strong reliability. Reliability estimation was also carried out using the results of the Item Response Theory (IRT) test through the item information function, as shown in **Figure 5**.


p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://dx.doi.org/10.20961/ijpte.v8i1.89913

**Table 1.** Confirmatory Factor Analysis

| Absolute Fit Measures | Description |
|---|---|
| The probable chi-square statistic ($\chi^2$) | Probability of $\chi^2_{count}$ = 0.385<br>The chi-square probability value, greater than or equal to, 0,05shows that the mode fits toward the data (good fit). |
| The Goodness of fit index (GFI) | In data analysis through the Lisrel program, the Goodness of Fit Index (GFI) = 0.91 is generated. Value GFI $\geq 0.90$ shows the model is suitable/fit to the data (good fit) |
| Adjusted goodness of fit index (AGFI) | The results of the data analysis show Adjusted Goodness of Fit Index (AGFI) = 0.85. Value AGFI $\geq 0.85$ shows heodel is suitable/fit to the data (good fit). |
| Root mean square residual (RMR) | Analysis of this data resulted in Root Mean Square Residual (RMR) = 0.015. This means that the model's ability can match the data well because the value of RMR is lesbian, or equal too, $0,05$ , shows the model is suitable/fit to the data (good fit). |
| Standardized RMR (SRMR) | Data analysis shows Standardized RMR = 0.011. Value of SRMR $\leq$ $0,05$demonstrates good modeling in terms of match/fitting the data (good fit). |
| Root mean square error of approximation (RMSEA) | Root Mean Square Error of Approximation (RMSEA) = 0.040. A value of RMSEA less than or equal to 0, 05 shows a good model ability in terms of match/fit data (good fit). |
| Expected Cross Validation Index (ECVI) | Data analysis shows the Expected Cross-Validation Index (ECVI) = 0.98. The ECVI value is used to compare several models. An ECVI value close to 1 indicates a model that better matches the data than other models. |

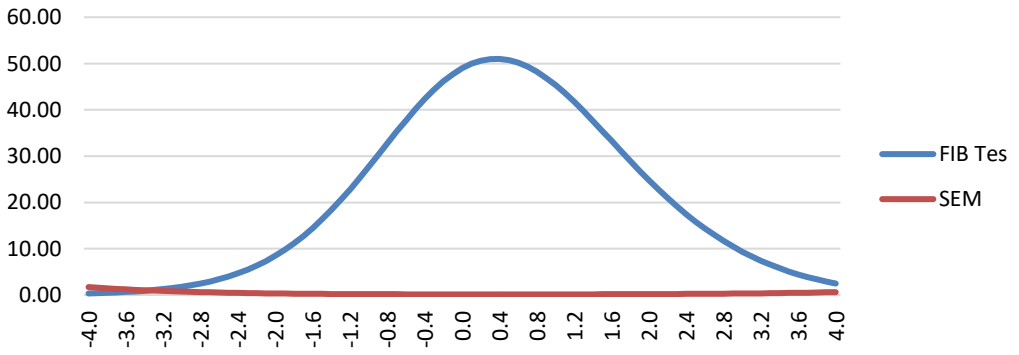


**Figure 5.** Item Information Function

The graph above depicts the item information function, with the X-axis representing the respondent's ability level and the Y-axis representing the size of the information function (Magis, 2013). The item provides significant information when given to subjects matching the examinee's ability level. This means the reasoning test instrument presents accurate information about the examinees, such as foreign speakers, explaining the reliability estimation. Implementing the item response theory model accurately measures the examinees' competence. IRT-based tests effectively explain the examinees' consistency (Istiyono et al., 2020).

Strict selection processes require a test with good firmness, as shown by the information function indicating measurement reliability. The Rasch model emphasizes the separation coefficient (item separation). Higher peaks of information signify greater measurement reliability. After proving validity and estimating reliability, the IRT assumption test included unidimensionality, parameter invariance, and local independence (Hambleton & Swaminathan, 1991). Dimensional factors with eigenvalues form each dimension in the analyzed instrument.


p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://dx.doi.org/10.20961/ijpte.v8i1.89913

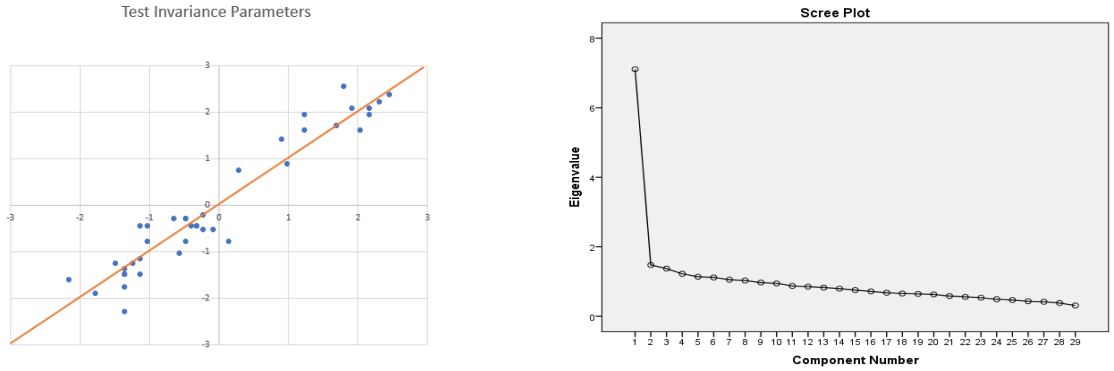


**Figure 6.** Unidimensional and Parameter Invariance Reasoning Test

Based on Figure 6. the unidimensional criteria are evident from the scree plot formed during the distribution analysis of examinees' abilities in answering each instrument. A test is unidimensional if components 1 and 2 in the scree plot have a long, steep gap (Furr, 2011). The analysis indicates that the first factor provides the highest contribution, more than double the second factor. The cumulative percentage of the three components, at 54%, sufficiently explains the test dimensions. The invariance test shows that each point is relatively close to the x=y diagonal axis, indicating no variation in parameter estimates for odd and even groups.

The ready-to-use instruments are then tested to determine the items' difficulty level, which is input to the CAT program.

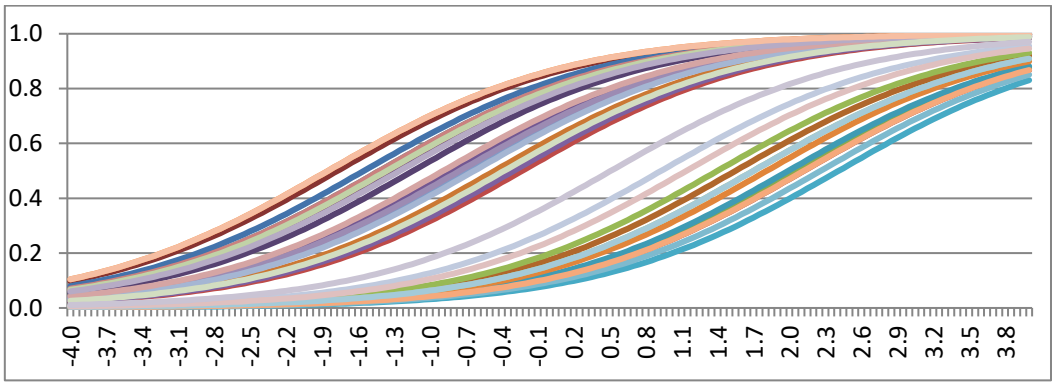


**Figure 7.** Difficulty Level of The Test Items

**Figure 7.** examinees' ability is measured using a logistic parameter, such as item difficulty level, with Rasch assisted by R Program version 2. The item difficulty level informs CAT development by selecting test items accepted by students. This conversion process uses Hambleton's theory, where abilities generally lie in the interval of -4 to 4 (Hambleton & Swaminathan, 1991). The conversion formula from ability -4 to 4 to a score of 0 to 100 is as follows:

$$Skor = \hat{\theta}\frac{100}{t + |r|} + \frac{100}{t + |r|}|r|$$

Explanation:
$\hat{\theta}$ = Ability estimation
t = The highest ability conversion (in this case, 4)
r = The lowest ability conversion (in this case, -4) (Hambleton & Swaminathan, 1991)


p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://dx.doi.org/10.20961/ijpte.v8i1.89913

After conversion, the examinees' abilities range from 0 to 100. This ability range reflects the high level of language reasoning students possess, including understanding the concept of facts, applying linguistic rules, breaking down information into specific forms, judging the value of ideas, combining word selection, idea formation, and context, and analogy and comparative thinking.

***Computerized Adaptive Test***

The development of the computerized adaptive test system resulted in a web application consisting of two portals: the admin portal and the participant (examinee) portal. This assessment method leverages technology to select and present items based on the estimated ability level of the examinee. CAT has been utilized for measurements across various fields. The display of the admin portal system is shown in **Figure 8**.

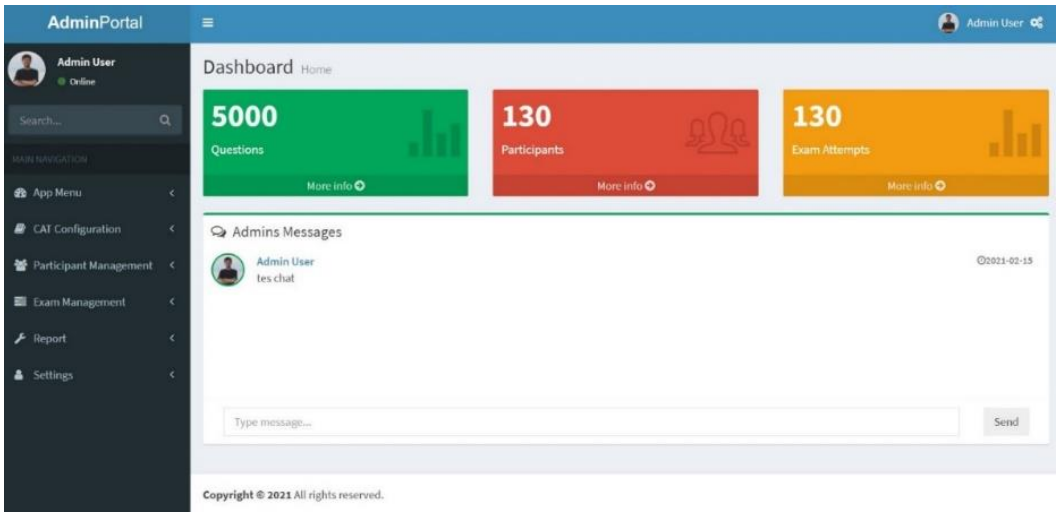


**Figure 8.** Admin System's View

**Figure 8** illustrates the system view on the admin portal, which includes several menus allowing admins to manage system data, such as question management, answers, exam result reports, and user data management, as shown in **Figure 9**.

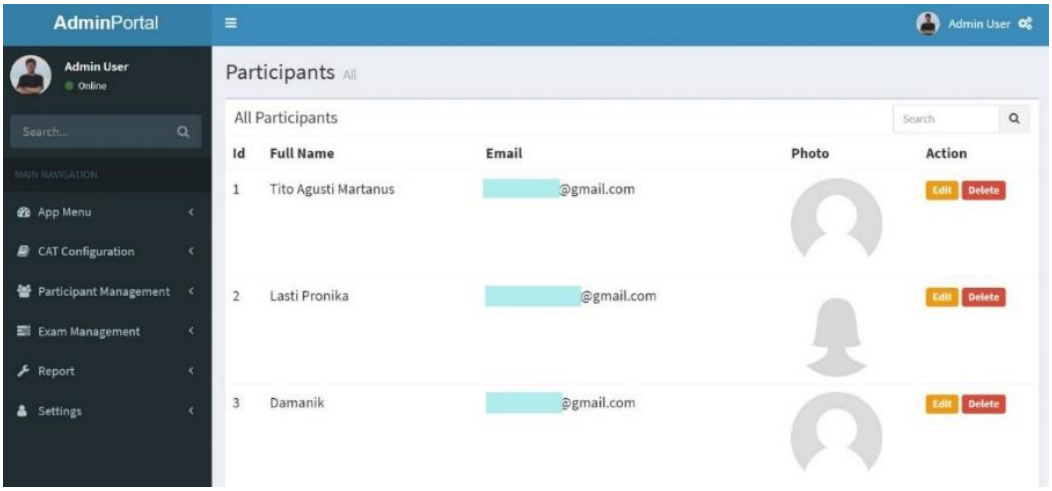


**Figure 9.** User Management on Admin's Portal


p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://dx.doi.org/10.20961/ijpte.v8i1.89913

The participant/examinee portal has two main sections: the dashboard for viewing available exams and test results and the display section for taking exams. The dashboard display of the participant/examinee portal is shown in **Figure 10**.

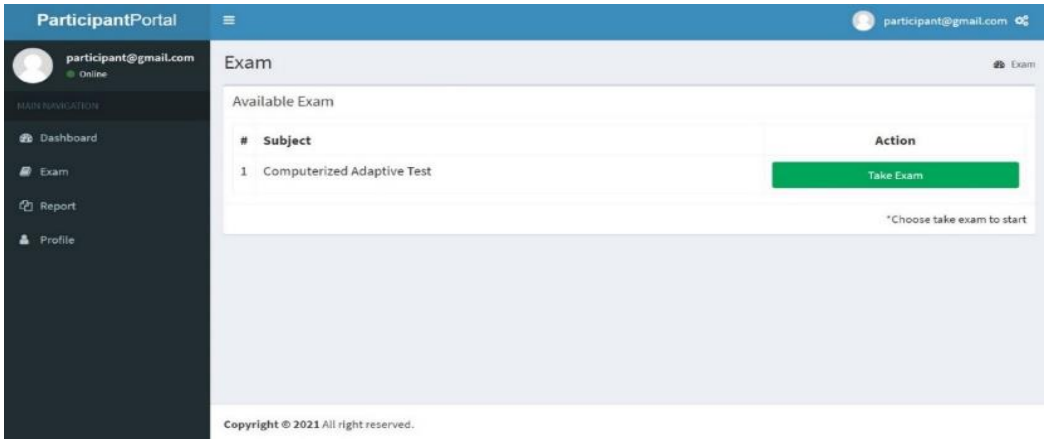


**Figure 10.** Participant's Portal View

**Figure 10** shows when a participant has logged in, with available menus such as profile updates, exams, and results. During the exam, the display guides the participant through the questions, as shown in **Figures 11** and **12**.

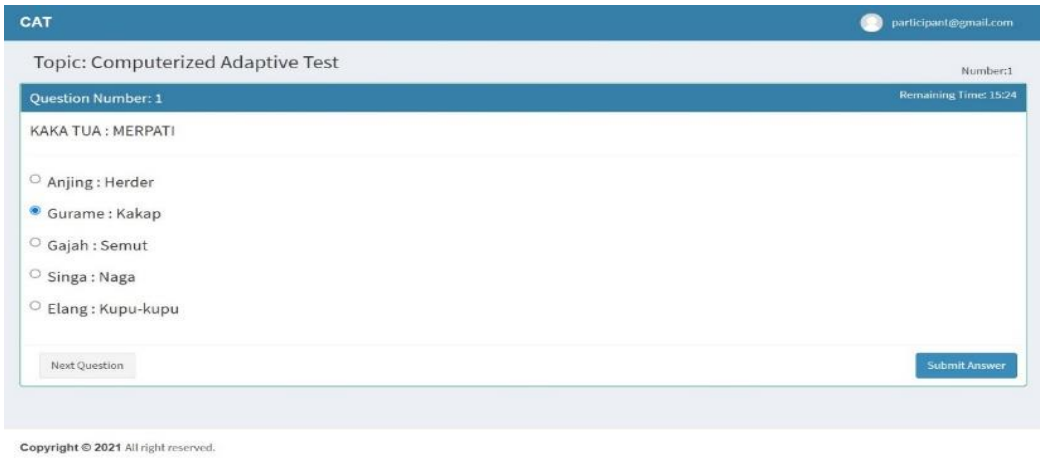


**Figure 11.** Page of Matching Words Exam

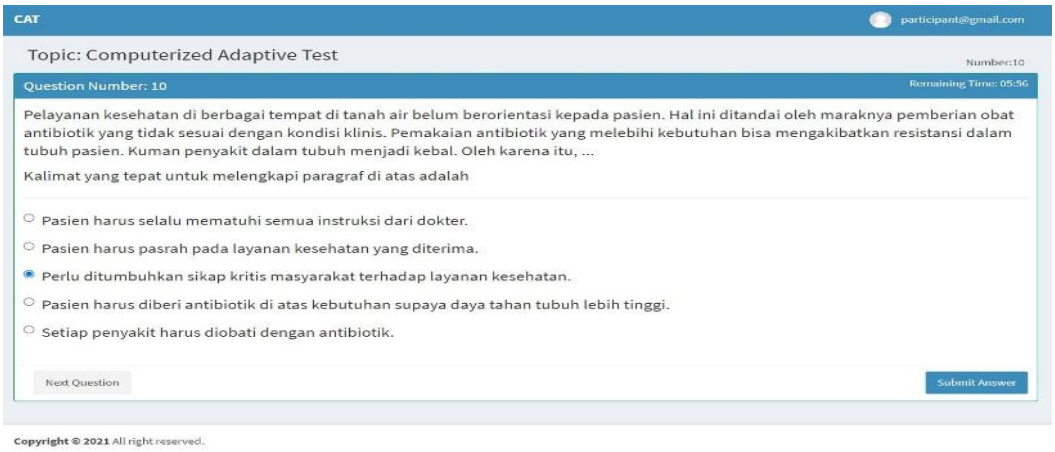


**Figure 12.** Page of Completing Paragraph Exam


p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://dx.doi.org/10.20961/ijpte.v8i1.89913

Figures 11 and 12 display the interface during the exam process, including the remaining time information, question display, multiple-choice answer options, save answer button, and submit button. Each time a participant saves an answer and proceeds to the next question, the adaptive test algorithm runs until the exam is complete. The user can see the test score upon completion, as shown in Figure 13.

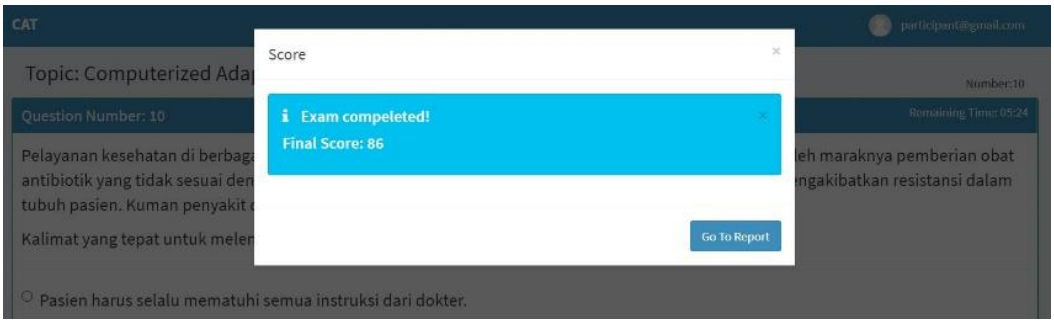


**Figure 13.** Popup Score Displayed Immediately after the Exam is Finished

Figure 13 shows a popup display of the exam results immediately after completion, with a 'Go to Report' button that leads to the report section in the participant portal, as shown in Figure 14.

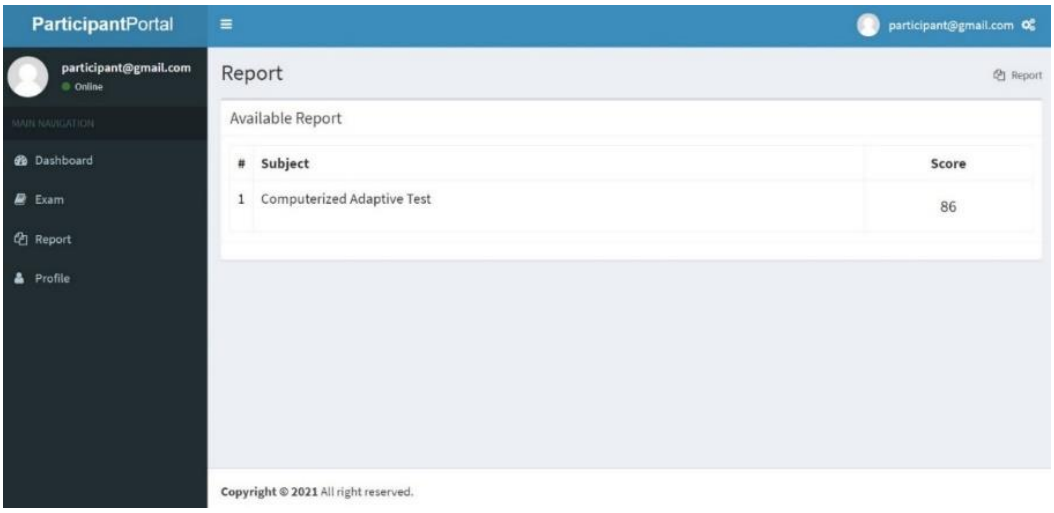


**Figure 14.** Display on Reports from Exams That Have Been Done

When the CAT starts, no questions are initially given to the participants, and no responses are available to estimate their ability level. Despite the lack of prior information about participants' abilities, the CAT implementation must begin. Suppose there is no initial information about the participants' abilities. In that case, the CAT starts by selecting an initial item corresponding to the moderate ability level at the item difficulty level [-1.00] to [0.100], ensuring all participants work on the same first item.

***Measurement of Language Reasoning Ability using CAT***

Determining students' reasoning competence in the social-humanities and science-technology fields involves obtaining rational considerations from experts who understand the need for tests and assessments for setting standards, comprehend the meaning of scores at various levels on the scale used to infer participants' performance, and fully grasp the limitations on capabilities associated with the required performance standards. Reasoning ability encompasses recognizing basic language forms: understanding the concept of facts, applying linguistic rules according to the agreements formed in the clauses read, breaking down information into more specific forms, judging the value of ideas, combining word selection, idea formation, and context, analogy


p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://dx.doi.org/10.20961/ijpte.v8i1.89913

thinking, and comparative thinking. Students' language reasoning ability is measured through these sub-competencies, presented in multiple-choice format. The CAT adjusts question difficulty based on the student's responses, presenting easier questions for incorrect answers and more difficult questions for correct answers. Based on CAT results from 300 participants, student abilities in the reasoning test are classified into low, medium, and high abilities (Retnawati et al., 2017).

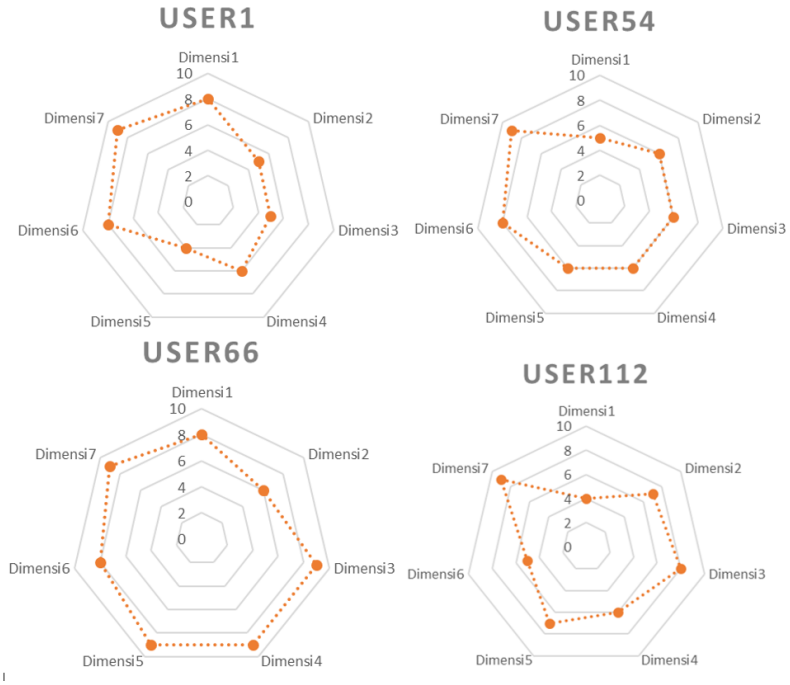


**Figure 15.** Map Sample of Student's Language Reasoning Ability

**Figure 15** shows each participant's access to the seven competency dimensions. Each dimension consists of validated questions representing the measured construct. Each user (test participant) knows their potential and reasoning ability after completing a series of CAT tests. Sumintono and Widhiarso explain that each measurement provides information about the results, focusing on the relationship between the test and the individual (Perera et al., 2018; Sumintono, 2018).

Understanding the concept of facts involves distinguishing discrepancies in the background or conclusions from presented fragments. Facts are states or events occurring in real life whose truth is proven and verified. Examinees must identify the suitability of the content, main information, and monologue parts that do not answer the question.

Applying language rules involves explaining whether the words used reflect the ideas they represent, whether the words are used in the correct context, whether the chosen words follow custom, and whether the selected words are used consistently. Based on the stimulus, test-takers must create structured, meaningful sentences that interact with the reader and represent effective writing (Rohmadi et al., 2020).

Reasoning ability also measures the participants' ability to identify paragraph unity, requiring each paragraph to have one topic and main idea. Paragraph development must support the topic without unrelated elements, ensuring effective communication. Ideally, one paragraph contains only one main idea. All sentences in a paragraph should address this main idea.

Examinees must also identify how to compare or contrast ideas in the question instrument to clarify explanations. Comparing or contrasting involves presenting similarities and differences between two things at the same level, both in terms of similarities and differences.


p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://dx.doi.org/10.20961/ijpte.v8i1.89913

***Effectiveness of CAT in Assessing Language Reasoning Skills***

The effectiveness of CAT in assessing language reasoning skills has been a focal point of this study. CAT's adaptive nature allows for tailored assessments based on individual performance, adjusting question difficulty dynamically. This capability ensures that each student is appropriately challenged, providing a more accurate measurement of their language reasoning abilities. By presenting questions that align with the examinee's skill level, CAT minimizes test fatigue and maximizes engagement, potentially yielding more reliable results compared to traditional static tests.

CAT's ability to provide immediate feedback enhances its utility in assessing language reasoning. The adaptive algorithm adjusts question difficulty and refines its estimation of the examinee's proficiency level with each response. This iterative process offers valuable insights into the nuanced aspects of language comprehension and reasoning, which are crucial in educational and professional settings.

***Implications for Educational Practice***

Findings from this study have significant implications for educational practice, particularly in curriculum design and assessment strategies. Employing CAT allows educators to gain deeper insights into students' language reasoning abilities beyond conventional standardized tests. This methodological shift promotes personalized learning experiences tailored to individual strengths and weaknesses, fostering a more supportive educational environment.

Integrating CAT into educational assessments can enhance the validity and reliability of evaluations in language-intensive disciplines. The adaptive nature of CAT ensures that students are adequately challenged, promoting a deeper understanding and application of language reasoning skills. This approach aligns with contemporary educational practices and prepares students more effectively for real-world challenges where logical thinking and clear communication are essential.

## 4. CONCLUSION

The ability to think logically represents good and accurate language, essential for identifying students' language competency in Indonesia. Based on this concept, this study developed two measurement products: standardized tests for Indonesian language reasoning skills for students in science, technology, and social-humanities fields, and a CAT categorizing students' reasoning abilities into low, medium, and excellent levels. The development of the CAT was conducted using an item response theory model, which can describe the probability of answering questions correctly based on the examinees' ability level and the item's difficulty. This study employed the one-parameter logistic (1PL) model, which uses the item difficulty parameter to determine the position of items in the item bank. The question bank (item bank) consists of calibrated items with established parameter values. An item bank was developed from academic potential test questions in the language field. CAT selects items from this question bank for test takers, and the availability of high-quality items in the question bank determines the quality of the CAT. The questions were derived from academic potential tests in the language field for university entrance. After identifying the difficulty levels of the test items, the test instrument was included in the question bank and re-tested on students in social-humanities and science-technology fields. The dimensions of the questions developed include understanding the concept of facts, applying language rules according to the agreements formed in the clauses read, breaking down information into more specific forms, judging the value of ideas, combining word selection, idea formation, and context, as well as analogy and comparison thinking. The proposed CAT system has proven effective in speeding up testing time and allowing students to complete the test according to their abilities. Computer-based assessment can incorporate multimedia to deliver better ideas and discourses, enhancing the assessment process for participants.


p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://dx.doi.org/10.20961/ijpte.v8i1.89913

## 5. REFERENCES


Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement, 45*(1), 131–142. https://doi.org/10.1177/0013164485451012

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford Publications.

Brown, T. A., & Moore, M. T. (2012). Confirmatory factor analysis. *Handbook of Structural Equation Modeling*, pp. 361–379. https://psycnet.apa.org/record/2012-16551-022

Clark, S. S. (2007). Thinking locally, suing globally: The international frontiers of mass tort litigation in Australia. *Defense Counsel Journal, 74*, 139.

Dove, G. (2014). Thinking in words: Language as an embodied medium of thought. *Topics in Cognitive Science, 6*(3), 371–389. https://doi.org/10.1111/tops.12102

Furr, R. M. (2011). Evaluating psychometric properties: Dimensionality and reliability. *Scale Construction and Psychometrics for Social and Personality Psychology*, pp. 25–51. https://dx.doi.org/10.4135/9781446287866.n4

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*(4), 347–360. https://doi.org/10.1111/j.1745-3984.1984.tb01039.x

Greenspan, S. I., & Shanker, S. (2009). *The first idea: How symbols, language, and intelligence evolved from our primate ancestors to modern humans*. Da Capo Press. https://doi.org/10.1093/brain/awh564

Hambleton, R. K., & Swaminathan, H. (1991). *Item response theory (Principles and applications)* (p. 107). Springer Science+Business Media. https://www.springer.com/gp/book/9780898380651

Istiyono, E., Dwandaru, W. S. B., Setiawan, R., & Megawati, I. (2020). Developing of computerized adaptive testing to measure physics higher order thinking skills of senior high school students and its feasibility of use. *European Journal of Educational Research, 9*(1), 91–101. https://doi.org/10.12973/eu-jer.9.1.91

Kavanagh, J. M., & Szweda, C. (2017). A crisis in competency: The strategic and ethical imperative to assessing new graduate nurses' clinical reasoning. *Nursing Education Perspectives, 38*(2), 57–62. https://doi.org/10.1097/01.NEP.0000000000000112

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*(4), 359–375. https://doi.org/10.1207/s15324818ame0204_6

Lane, S., Raymond, M. R., & Haladyna, T. M. (2015). *Handbook of test development*. Routledge. https://doi.org/10.4324/9780203102961

Larson, J. W., & Madsen, H. S. (2013). Computerized adaptive language testing: Moving beyond computer-assisted testing. *Calico Journal, 2*(3), 32–37. https://eric.ed.gov/?id=EJ340118

Leighton, J. P. (2006). Teaching and assessing deductive reasoning skills. *The Journal of Experimental Education, 74*(2), 107–136. https://eric.ed.gov/?id=EJ340118

Magis, D. (2013). A note on the item information function of the four-parameter logistic model. *Applied Psychological Measurement, 37*(4), 304–315. https://doi.org/10.1177/0146621613475471

Perera, C. J., Sumintono, B., & Jiang, N. (2018). The psychometric validation of the Principal Practices Questionnaire based on item response theory. *International Online Journal of Educational Leadership, 2*(1), 21–38. https://doi.org/10.22452/iojel.vol2no1.3

Ramadhan, S., Sumiharsono, R., Mardapi, D., & Prasetyo, Z. K. (2020). The quality of test instruments constructed by teachers in Bima Regency, Indonesia: Document analysis. *International Journal of Instruction, 13*(2), 507–518. https://doi.org/10.29333/iji.2020.13235a

Retnawati, H. (2015). The comparison of accuracy scores on the paper and pencil testing vs. computer-based testing. *Turkish Online Journal of Educational Technology-TOJET, 14*(4), 135–142. https://eric.ed.gov/?id=EJ1077660



**p-ISSN: 2597-7792 / e-ISSN: 2549-8525**
**DOI:** https://dx.doi.org/10.20961/ijpte.v8i1.89913

Retnawati, H., Hadi, S., Nugraha, A. C., Arlinwibowo, J., Sulistyaningsih, E., Djidu, H., & Apino, E. (2017). Implementing the computer-based national examination in Indonesian schools: The challenges and strategies. *Problems of Education in the 21st Century, 75*(6), 612. https://doi.org/10.33225/pec/17.75.612

Rohmadi, M., Ulya, C., Sudaryanto, M., & Ximenes, M. (2020). Feasibility analysis of basic writing and reading materials for foreign speakers. *Proceedings of the 2nd Konferensi BIPA Tahunan by Postgraduate Program of Javanese Literature and Language Education in Collaboration with Association of Indonesian Language and Literature Lecturers, Surakarta: 9 November 2019*. https://doi.org/10.4108/eai.9-11-2019.2295063

Ryle, A. (2012). Critique of CBT and CAT. *Change for the Better, pp. 4*, 1–8. https://in.sagepub.com/sites/default/files/upm-binaries/47043_07_Critique_of_CBTand_CAT.pdf

Sudaryanto, M., Mardapi, D., & Hadi, S. (2019a). Multimedia-based online test on Indonesian language receptive skills development. *Journal of Physics: Conference Series, 1339*(1), 1–7. https://doi.org/10.1088/1742-6596/1339/1/012120

Sudaryanto, M., Mardapi, D., & Hadi, S. (2019b). How foreign speakers implement their strategies to listen Indonesian language? *Journal of Advanced Research in Dynamical and Control Systems, 11*(7), 355–361. https://www.jardcs.org/abstract.php?id=2551#

Sudaryanto, M., Ulya, C., Rohmadi, M., & Kuhafeesah, K. (2020). Inter-rater assessment on listening media for foreign language speakers. *Proceedings of the 2nd Konferensi BIPA Tahunan by Postgraduate Program of Javanese Literature and Language Education in Collaboration with Association of Indonesian Language and Literature Lecturers, Surakarta: 9 November 2019*. https://doi.org/10.4108/eai.9-11-2019.2295064

Sumintono, B. (2018). Rasch model measurements as tools in assessment for learning. *Advances in Social Science, Education and Humanities Research, 17*(3), 38–42. https://doi.org/10.2991/icei-17.2018.11

Triyono, Senam, Jumadi, & Wilujeng, I. (2017). The effects of creative problem solving-based learning towards students' creativities. *Jurnal Kependidikan: Penelitian Inovasi Pembelajaran, 1*(2), 214–226. https://doi.org/10.21831/jk.v1i2.9429

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. Routledge. https://doi.org/10.4324/9781410605931

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*(4), 361–375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x


**p-ISSN: 2597-7792 / e-ISSN: 2549-8525**
**DOI:** https://dx.doi.org/10.20961/ijpte.v8i1.89913