The development of computer adaptive test and item response theory with 4 parameters based the logistics model

L S Riza*, N S Nurjanah and Y Wihardi
Department of Computer Science Education, Universitas Pendidikan Indonesia, Indonesia

*Corresponding author’s email: lala.s.riza@upi.edu

Abstract. Students’ abilities can be measured by making a test whose the number and content of questions have been determined for each student. However, it is less effective because the distribution of students' abilities is different. By using Computer Adaptive Test (CAT) and Item Response Theory (IRT), the test can be done more quickly even though each student might solve different number and content questions. Therefore, in this research we develop CAT and IRT with 4 parameters based logistics model to evaluate the students’ ability. Some experiments conducted on 27 students with 50 number of questions were analyzed and compared with Classical Test Theory (CTT). The results indicate the system can be used as an alternative tool for evaluating the students’ ability.

1. Introduction
There are three important things in an educational evaluation namely, assessment, measurement and measuring tools used. In an assessment requires measurement, and in a measurement requires a measuring instrument. The assessment is a qualitative description of students' behavior both based on the measurement result and not the measurement result [1], while the measurement is the score determined based on the results of the test [2]. The test is a measuring instrument used in measurement and assessment in the evaluation of education.

Presentation of questions to test participants is also important. The most commonly used method today is paper and pencil test but with technological development, the presentation of technologically assisted test also developed, such as computerized test (CBT) and computerized adaptive test (CAT). CBT focuses on changing the paper and test method into the computer version [3]. Meanwhile the CAT is a computer-based test where the items given to the test participants are tailored to the testers' abilities [4].

After the test or measurement process conducted by using the test is completed, then the next is to analyze the results of these measurements. There are two ways to analyze test results, namely classical test theory (CTT) or classical test theory and item response theory (IRT) or modern test theory. CTT is a theory of testing based on the idea that the value observed or obtained by a person on a test is the number of true scores (error free score) and error score [5]. Although the name is a classic test but this theory is still widely used by educators. While modern test theory developed based on classical test theory. In the IRT, the true student response opportunities, grammatical characteristics or parameters, and the characteristics or parameters of test participants are linked through a model formula that must be adhered to by either the test group or the test group [6]. The disadvantages of classical test theory are that there is a link between a group of students and a grain group or a particular package of questions,
then modern test theory eliminates that link. Any matter can be given to any test participant, and vice versa and can still be calculated ability.

In the IRT there are two parameters used in analyzing the test ie, the ability parameter is also called theta (θ) and the test parameter. In IRT there is a logistic model consisting of 4 logistic models i.e., 1 logistic parameter (1PL) which has 1 parameter i.e., difficulty level, 2-parameter logistic (2PL) which has 2 parameters i.e. different power and difficulty level, 3-parameter logistics (3PL) which has 3 parameters i.e., different power, level of difficulty and guess factor and 4-parameter logistics (4PL) [6]. The four models are based on the number of parameters used. The parameters used in IRT 4PL namely, differentiation, difficulty level, guess factor and carelessness. 3 other factors have been contained in IRT 4 PL and the 4th factor of carelessness is a factor that does not exist from the other three models. Carelessness factor or carelessness factor is a factor that makes the student with high ability may answer wrongly on problem which has difficulty level below its ability [7].

Based on the problems and motivations mentioned above, we focus on implementing tests in the form of computerized adaptive testing based on 4-parameter response theory items. The CAT development is done on the grounds that CAT has many advantages in addition to being able to cover the shortcomings of paper and pencil test and CBT but also able to conduct direct and accurate assessment, individual timed tests and administrative tests can be repeated, standard instruction and administrative conditions during the exam, security better test tests (not available in physical form) and available test test presentations for participants with special needs tests [8]. Meanwhile, IRT is suitable for analyzing test results performed using CAT. Because of the adaptive nature of each participant get different items of question, but still can be calculated ability level using IRT.

2. Methods

2.1. Research Design

Based on Figure 1 can be described the research process as follows:

2.1.1. The preparation phase is the initial stage of this research. At this stage it begins by identifying the problem first, then searching for relevant theories and methods to get the settlement of the problem. Then in the final stage is to decide the theories and methods to be used for research.

2.1.2. The literature study is conducted as part of the preparation phase. At this stage is done by searching, studying and understanding the theories and methods used to conduct research. Some theories that need to be understood to be able to do this research, among others, the theory to determine the ability of students that is item response theory (IRT) with 4 parameters logistics and theory of how to present the problem in computerized adaptive testing (CAT). The theories are studied based on the literature that has been collected through journal, books, and articles related to the research.

2.1.3. Data collection. At this stage is done by collecting the source of the problems obtained from a basic programming book used in the learning process in school as many as 50 questions. Then the questions that have been obtained are used to test the initial problem to get the initial value of the 4 parameters and the probability of the correct answer which will be used as the feature of the problem. 4 parameters such as power difference, level of difficulty, guess factors and factors of ignorance. After obtained the answers from 50 students the next question is calculated the different power, the level of difficulty and guess factors based on the student's answer, while for the fourth factor obtained from previous literature studies. Which then will all be a question bank.

2.1.4. Development on CAT with IRL 4 parameter. At this stage is done by considering the waterfall model, involving analysis on user requirement, design, coding, and testing.
2.1.5. Experiments. After the software development is done, the next step is to conduct an experiment. Experiments were conducted on vocational students who had previously obtained the materials before.

2.1.6. Analysis and Conclusion. After the experiment is completed next is to analyze whether the application of computerized adaptive testing (CAT) with Item Response theory theory (IRT) 4-parameters logistics can really describe the ability of test participants well.

![Research Design on Developing CAT with IRT 4 Parameters](image)

**Figure 1.** Research Design on Developing CAT with IRT 4 Parameters.

2.2. Development of CAT with IRT 4PL

Figure 2 shows the procedures on CAT with IRT 4PL as follows:

2.2.1. Creating item sets. At this stage, questions are compiled based on data collection in the previous sub-chapter. The question data is structured in such a way as to fit the material taught in the school at that level.

2.2.2. Pilot testing. This stage is for checking, evaluating and calculating initial parameters of IRL.

2.2.3. Calculating IRL parameter. After conducting pilot testing, we calculate the IRL parameters related to different power, difficulty level, factor of guessing, and carelessness factor. The four parameters are later used as the basis in determining the selection of questions that will be given to participants and calculations to determine the ability of participants.

2.2.4. Real testing using the proposed system. The question data that have been completed with these parameters are inserted into the question bank within the proposed application.
2.2.5. Students’ answers. First, the application provides questions to participants with moderate difficulty levels. Then, the application takes students’ responses to the problem in the form of right or wrong answers.

2.2.6. Calculating parameters. The next step, the application to calculate based on the parameters that exist on the given problem. The calculation begins with calculating $\theta$ (the participant's ability), $P(\theta)$ (the probability of the participant to answer correctly), $I(\theta)$ (the answerable information function), $SE$ (the standard error by application), and $SE$ margin.

2.2.7. Stopping Rule. Then, the application examines the given criteria of termination. There are three criteria, as follows: the items are not available any more, the time of testing is over, and the difference of standard error has less than 0.01. If the previous problem can be answered correctly then the next problem is a matter that has a greater degree of difficulty. While if one then the given is a matter that has a level of difficulty is lower than before.

2.2.8. Students’ final remarks. If the stopping rule is met, then the final ability of the participants is obtained. The final ability of the participants can be used by the teacher in accordance with the goals and needs.

![Diagram](image)

**Figure 2.** The procedures on CAT with IRL 4PL

2.3. Experimental Design

To conduct this research activity required samples, samples taken from the population having certain qualities and characteristics. In this study the population is students of SMK Negeri 2 Bandung. Sample determination technique used in this research is purposive sampling technique that is technique used in choosing sample based on certain consideration. The selected sample is a sample that meets the criteria required in this study, that are 27 students in eleventh class.

3. Results and Discussion

The experimental results for all participants can be seen in Table 1. All participants have a moderate degree of competency level from a scale of -3 to 3. The lowest ability level is participants 18 and participants 19 who have a capability level of -0.0581689. Meanwhile, the highest level of ability is
0.463351 obtained by participants 9. This shows the level of ability of participants spread with a difference that is not too significant.

Table 1. Experimental Results Using CAT.

| No | Participant Names | Final $\theta$ | No | Participant Names | Final $\theta$ |
|----|-------------------|----------------|----|-------------------|----------------|
| 1  | P1                | -0.0410906     | 15 | P15               | 0.204611      |
| 2  | P2                | 0.0774592      | 16 | P16               | 0.0478528     |
| 3  | P3                | -0.00179233    | 17 | P17               | -0.0391514    |
| 4  | P4                | -0.107695      | 18 | P18               | -0.0581689    |
| 5  | P5                | -0.10567       | 19 | P19               | -0.0581689    |
| 6  | P6                | -0.110194      | 20 | P20               | -0.109299     |
| 7  | P7                | -0.0410906     | 21 | P21               | 0.0482633     |
| 8  | P8                | 0.0482633      | 22 | P22               | 0.0176115     |
| 9  | P9                | 0.463351       | 23 | P23               | 0.0473627     |
| 10 | P10               | 0.0239316      | 24 | P24               | 0.0516663     |
| 11 | P11               | -0.110194      | 25 | P25               | 0.046508      |
| 12 | P12               | 0.0516663      | 26 | P26               | -0.0669112    |
| 13 | P13               | 0.0478528      | 27 | P27               | -0.109299     |
| 14 | P14               | -0.109706      |    |                   |               |

The analysis of the results of the experiments performed using IRT-4 logistic parameters when compared with the analysis of the capability of the participants of experiments by using classical test theory shows the differences that can be seen in Table 2. The value of capability with IRT 4-parameters logistic analysis and classical test theory on participants 1 shows a difference of 0.062404, the value does not show the excessive difference of the two forms of analysis of the test results. Meanwhile, the value of ability with IRT 4-parameters logistics analysis and classical test theory on participants 2, 3, 4, 15, 26 showed a considerable difference that is above 0.1. This is due to the ability to calculate the capabilities in the two different types of test results analysis. In the classical test analysis each question is given the same weight, whereas in the IRT-4 analysis the logistics parameters take into account the attributes attached to the problem. So that the weight given to each problem will vary depending on the value of different power, the level of difficulty, the guess factor and the factor of carelessness attached to the problem.
4. Conclusion
This research has been developing a system used for assessing students’ ability by using CAT with IRL 4 parameters. The development involve several stages, as follows: creating items, pilot testing, calculating IRL parameter, real testing using the proposed system, students’ answer, calculating parameters, stopping rules, and calculating students’ final remarks. An experiment to validate has been conducted as well. Moreover, a result comparison between the proposed system using CAT with IRL 4 parameters with the classical test theory has been presented. In the future, we consider and combine other methods for educational assessment as conducted in [9,10].

5. References
[1] Norman G 1982 Constructing Achievement Tes New York: Prentice Hall-Ihc
[2] Phillips DA & Hornak JE 1979 Measurement and Evaluation in Physical Education Wiley
[3] Scissons E H 1976 Computer administration of the California Psychological Inventory Measurement and Evaluation in Guidance 9 1 22-25
[4] ÖZYürt Ö, ÖZYürt H and Baki A 2013 Design and development of an innovative individualized adaptive and intelligent e-learning system for teaching–learning of probability unit: Details of UZWEBMAT Expert Systems with Applications 40 8 2914-2940
[5] Novick M R 1966 The axioms and principal results of classical test theory. Journal of mathematical psychology 3 1 1-18
[6] Hambleton R K and Swaminathan H 2013 Item response theory: Principles and applications Berlin: Springer Science & Business Media
[7] Magis D 2013 A note on the item information function of the four-parameter logistic model. Applied Psychological Measurement 37 4 304-315.
[8] Segall D O 2005 Computerized adaptive testing Encyclopedia of social measurement 1 429-438
[9] Farasyi G, Setiawan W, Fahsi M and Riza L S 2018 The Association Rule Method for Mapping and Recommendation System on Students’ Difficulties. Transylvanian Review 1 1
[10] Riza LS, Awaludin R, Sutarno H, Munir and Wibawa A P 2017 A model for auto generating sets of examination items in educational assessment by using fuzzy c-means World Transactions on Engineering and Technology Education 15 2 114-119.