Analysing PISA-like assessment test measuring Scientific Literacy using Three-Parameter Logistic (3PL) of IRT—2018

S Rohmah1*, I Kaniawati2 and T R Ramalis2

1Student of Postgraduate School, Universitas Pendidikan Indonesia, Indonesia
2Postgraduate School, Universitas Pendidikan Indonesia, Indonesia
*Corresponding Author. Email: siti.rohmah@student.upi.edu

Abstract. This study investigated the multiple-choice of scientific literacy resembling to PISA framework assessment test to assess senior high school students’ literacy acquisition in physics by applying Item Response Theory (IRT) with three-parameter logistic. The multiple-choice instrument composed by 15 items test measuring three aspects of the scientific literacy assessment in PISA 2015 framework such as Context, Knowledge and Competencies. Three-parameter logistic (3PL) used to analysing test are difficulty, discriminatory, and guessing parameters of the instrument items were fitted by IRT model, using EIRT. Data was gathered on 32 students in science program of student high school in Bandung region. The IRT framework reveals item-level information and indicates appropriate ability ranges for the test. With the result that these methods can apply to diagnose and evaluate the features of items at various scientific literacy aspects of test takers.

1. Introduction
The PISA-like assessment test developed by Maulana and Suhand at in 2016, is known standard multiple-choice test and Likert test of scientific literacy for physics course at high school student. The test consists of two types of test. The first type of test be made up of 15 items of multiple choices test that covers three domain of scientific literacy framework developed PISA which is context domain, knowledge domain, and competencies domain. The second type of test be made up of 18 statement that cover the fourth domain which is attitude domain [8]. All domains of scientific literacy are important to measure student’s proficiency to learn about physics as part of science. Previous research that using this test has been demonstrated classical test assessment methods: the item difficulty index and the item discriminatory index [1]. However, the framework of classical test theory (CTT) for test assessment has some important limitations. For instance, the item parameters depend on the ability distribution of examinees and the ability parameters depend on the set of test items. To overcome these shortcomings, the item response theory (IRT) was introduced [2–4].

Therefore, the purpose of this study is to explore PISA-like assessment test based on the framework of IRT. We will first present the key concepts of IRT, focusing on the three-parameter logistic (3PL) model used in the study (Sec. 2). Section 3 is the research method that used by the test from tenth grade high school students that study physics. Section 4 (results and discussion) will be discussed some limitations of CTT also advantages of IRT, and the significance of using 3PL-IRT analysed by our data. Last, in Sec. 5, we will summarise what we did and found in this study.

2. Item Response Theory
The IRT framework rests on the assumption that the performance of an examinee on a test item can be predicted from the item’s typical features and the examinee’s latent traits—often called abilities or
person parameters. The relationship between the examinees’ performance on an item and their ability is described by an item characteristic function (or item response function), which quantitates how the probability of a correct response to a specific item increases with the level of an examinee’s ability. The graph of this relationship is known as the item characteristic curve (ICC). The empirical ICCs in prior published research, of relevance to IRT, tend to be S-shaped or sigmoidal. As the ability increases, the empirical ICC rises slowly at first, more sharply in the middle, and again slowly at very high levels of ability. In its early days, the normal ogive function was commonly used to model the ICC, while nowadays the logistic function is a popular alternative, as shown in Eq. (1). An example of the ICC from our data, established by fitting a logistic function, is shown in Fig. 1.

\[
P_i(\theta) = c_i + (1 - c_i) \frac{1}{[1 + \exp(-Da_i(\theta - b_i))]} \tag{1}
\]

where \(b_i\) is the item difficulty parameter (\(\theta\) at the inflection point of the ICC), and \(a_i\) is the slope of the ICC at that inflection point, called the item discriminatory parameter. The lower asymptote level \(c_i\) of the ICC, which corresponds to the probability of a correct answer at very low-ability levels, is referred to as the pseudo-chance level or sometimes as the guessing parameter. The constant \(D\) normally equals 1.7, and is used to make the logistic function as close as possible to the normal ogive function. The 3PL model can be reduced to the two-parameter logistic (2PL) model by setting \(c = 0\). The 2PL model is most plausible for open ended questions, in which responses are rarely guesses. Moreover, this can be further reduced to a one-parameter logistic (1PL) model by considering only the \(b\) parameter at which the probability of a correct response is 0.5, while holding a fixed [2,7–8]. The special cases with \(D = 1.0, c = 0,\) and \(a = 1.0,\) are known as Rasch models [5].

In this study, we applied the EIRT: Item Response Theory applications using Excel add-in created by a collaborative effort between Université du Luxembourg and Université Laval (Québec, Canada). EIRT has integrated well-known models, namely, the 1PLM, 2PLM, and 3PLM to estimate the appropriate parameters. Also included in the program is a generalization of the logistic model to accommodate multiple-choice items. One possible generalization is Bock’s nominal model. This model is appropriate with nominal data, for instance, with a four-option multiple-choice item. EIRT also provides classical test theory statistics such as Cronbach’s alpha [6].
3. Data Collection
The 15-item PISA-like assessment test was validated by researcher in physics education that used to measure student’s scientific literacy in physics class with topic static fluid. The test was revised based on suggestions from the professors. We applied the assessment to 32 student high school in Bandung. These students had learned static fluid concepts through lectures and integrated demonstrations and group discussions by teacher. The participants took 30 min to complete the assessment. The collected data were analysed using the 3PL model in IRT and Instrument information function plots.

4. Result and Discussion
In applying IRT to the data gathered on tenth grade high school student in Bandung, we used the add-ins on Microsoft program called EIRT to fit the three-parameter logistic (3PL) models, one for each PISA-like assessment test item. We assumed a single ability, named the scientific literacy ability, which represents the latent traits in each student that affect performance in the scientific literacy. Each logistic model is determined by identifying its three parameters: discrimination \( a \), difficulty \( b \), and guessing \( c \). These identified parameters are shown in Table 1 for the 15 items test, categorised by physics concepts in static fluid. Moreover, the item difficulty index \( (diff) \) and discrimination using the point-biserial coefficient \( (r_{pb}) \) from the CTT framework are included as the last columns of Table 1.

The criterion ranges of the item parameters are shown by interval bounds in square brackets.

Table 1. The Model Parameters Identified in IRT Analysis, Namely, Discrimination \( a \), Difficulty \( b \), And Guessing \( c \), for the 15 Items in PISA-Like Assessment Test Categorized by Static Fluid Concepts, Along with the Item Difficulty Index \( (Diff) \) and the Discrimination \( (disc) \) from CTT Analysis

| Static Fluid Concept | Item | Discrimination \( IRT\ (a) \) | \( CTT\ (disc)b \) | Difficulty \( IRT\ (diff)c \) | \( CTT\ (diff)b \) \( \geq 0.18 \) | Guessing \( IRT\ (c) \) | \( r_{pb} \) |
|----------------------|------|------------------|------------------|------------------|------------------|------------------|--------|
| Pascal’s law         | 1    | 0.595            | 0.000            | 0.622            | 0.500            | 0.166            | 0.66   |
|                      | 5    | 0.623            | 0.093d           | 3.182            | 0.313            | 0.194            | 0.489  |
|                      | 8    | 0.628            | 0.164            | -0.435           | 0.625            | 0.168            | 0.225  |
|                      | 10   | 0.677            | 0.073            | 1.352            | 0.406            | 0.166            | 0.056  |
|                      | 12   | 0.719            | 0.062            | -1.398           | 0.750            | 0.167            | 0.124  |
| Archimedes principle | 3    | 1.005            | 0.239            | -0.542           | 0.656            | 0.164            | 0.750  |
|                      | 7    | 1.546            | 0.200            | 0.476            | 0.438            | 0.154            | 0.438  |
|                      | 11   | 0.884            | 0.056d           | 31.790d          | 0.188            | 0.180            | 0.167  |
|                      | 14   | 0.931            | 0.022d           | -1.637           | 0.813            | 0.167            | 0.124  |
| Capillarity          | 2    | 1.443            | 0.263            | -1.257           | 0.813            | 0.164            | 0.167  |
|                      | 4    | 3.275            | 0.274            | 0.631            | 0.281            | 0.124            | 0.167  |
|                      | 13   | 0.607            | 0.000            | 0.683            | 0.500            | 0.173            | 0.167  |
| Viscosity and        | 6    | 0.843            | 0.225d           | 3.966            | 0.188            | 0.167            | 0.167  |
| Stoke's law          | 9    | 0.214            | 0.489d           | -3.993           | 0.750            | 0.171            | 0.167  |
|                      | 15   | 0.773            | 0.015d           | 1.863            | 0.344            | 0.173            | 0.167  |

\( a \) in classical test theory, the item discrimination is the point-biserial correlation coefficient between the item scores (0 or 1) and the total test scores

\( b \) in classical test theory, the item difficulty is the percentage of correct response, so a low value indicates a high item difficulty

\( c \) in classical test theory, there is no formal definition of guessing chance. Guessing chance of a multiple-choice question is often estimated as \( (1/ \text{number of choices}) \), which is 20% for every item in five multiple choice tests. This estimation implies the probability of choosing each choice is equal.

\( d \) outside the criterion range.

The item difficulty \( b \) is the ability \( \theta \) at the inflection point of ICC. In the logistic model, at this point, the probability of correct answer is \( (1 + c)/2 \), midway between the asymptote levels, as seen by
substituting \( \theta = b \) in Eq. (1). When \( c = 0 \), as in the 1PL and 2PL models, the probability of a correct answer is 0.5 and this could be used to identify \( b \). Parameter \( b \) is named “difficulty” because a harder test item requires higher ability \( b \) for probability 0.5 of a correct answer. The criterion range for \( b \) is chosen to [-3,3]. Clearly, an item with \( b \) close to -3 is very easy, while \( b \) close to 3 is very difficult for the sample population of examinees.

The results in Table 1 show that item 6 was the most difficult question for the group of students, with the maximum \( b = 3.9 \). Item 6 has only 18% of the students correctly answered item 6 (diff = 0.18). Since only a few students from the high-ability group could answer the test correctly, item 6 clearly presented the most difficulty to the normal-ability students. Although difficulty \( b \) from the IRT analysis and the diff from the CTT analysis differ in their theoretical interpretations, they tend to be in good agreement. For example, item 5 (\( b = 3.1 \)) and item 15 (\( b = 1.8 \)) were somewhat difficult for the students, according to IRT. Their correct response rates were only 31% and 34%, respectively, indicating them as hard items according to the diff values also. Further agreement was found in item 14, whose CTT-diff indicated an easy question, with about 81% correct answers, consistent with difficulty \( b = -1.6 \). However, the two measures of difficulty seem to disagree on item 4. Viewed through CTT, item 4 appeared to be quite difficult items with correct answer rates 28%, but their respective \( b \) values were 0.6 quite easy for the difficulty parameter of IRT. As discussed earlier, this is one downside of the CTT approach, whose results depend on the examinees’ ability. The guessing parameter \( c \) of an item represents the probability that an examinee with very low ability level answers correctly. This may relate to the attractiveness of the answer choices and/or the guess behaviour of the examinees. Its value is equal to the level of a lower asymptote for the ICC, and it ranges from 0 to 1. Typically, \( c \) should be less than 0.3 [7]. Table 1 shows that, overall, the very low-ability students had less than a 30% chance of choosing the correct option for most the PISA-like assessment test items.

To show how the probability of correct response for a specific item depends on the ability of an examinee, we build the item characteristic curve (ICC). The ICC of a well-designed question should have a sigmoidal S shape [2,7]. Then the probability of a correct response would consistently increase with ability, and a high slope at the inflection point would indicate sharp separation by ability around that point. As shown by the solid line in Fig. 2, item 4 of the PISA-like assessment test mostly agrees with these criteria in our data. It has high discrimination power (\( a = 3.275 \)) for separating examinees at medium-ability level near \( b = 0.63 \), and very low-ability students have a 28% chance to correctly answer the item (\( c = 0.12 \)).

![Figure 2](image-url)  
**Figure 2.** The ICCs for items 2, 3, 4, and 7 in the PISA-like assessment test.

Overall, the results show that each PISA-like scientific literacy assessment test item has proper discrimination power at its specific difficulty (or ability level). Clearly, the steepness of the curve demonstrates the capability of the item to discriminate in the ability domain between examinees who understand the concept targeted by the item and those who do not. In general, a set of test items or questions should cover a range of ability domains in which the test takers are expected to differ. To determine how well the assessment does in testing adoption of the physics concept by the examinees at their various ability levels, the test information function was investigated in IRT framework.
The information function for a test at $\theta$, denoted $I(\theta)$, is defined as the sum of the item information functions at $\theta$:

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta) = \sum_{i=1}^{n} \frac{[P'_i(\theta)]^2}{P_i(\theta)Q_i(\theta)}$$  \hspace{1cm} (2)$$

where $I_i(\theta)$ is the item information function of item $i$, $P'(\theta)$ is the derivative of $P(\theta)$ with respect to $\theta$ for item $i$, and $Q(\theta) = 1 - P(\theta)$ [2,7]. This feature is not available in CTT. For example, the point biserial coefficient for an item is influenced by all items in the test. A plot of $I(\theta)$ for the 15-item PISA-like assessment test across ability levels of scientific literacy is shown in Fig. 3. The information curve peaks sharply to its maximum value at $\theta = 1.12$. This indicates that the test provides information about the scientific literacy ability of student most effectively when the examinees have abilities roughly in the range from 0.1 to 2.1 (medium to high). For cases with scientific literacy abilities less than 0, the PISA-like test provides very little information that would distinguish their differences, while the results of the 15-item test are highly sensitive to ability differences around ability 1.1. In general, the purpose of the test decides what type of information curve would be desired. For example, a flat curve over the whole range of abilities indicates that the component items are sensitive to variations at different ability levels, so the test information obtained by summing the item information becomes evenly distributed. This is desired if the test serves to assess a wide variety of abilities. In contrast, a sharply peaked curve sensitively reports differences of the test takers only around that peak, elsewhere it acts like a pass or fail threshold. This may be desirable when the purpose of the test is to provide eventual pass or fail labels. Test developers can benefit from these item and overall test information curves, to revise tests with such considerations of purpose in mind.

5. Conclusion

Results in this study indicate that the 15-item scientific literacy test, with 5 choices per item, is useful for testing students’ literacy acquisition in physics when the ability of the examinees is from medium to high. It can be applied as a pass or fail threshold instrument at a somewhat high-ability value ($\theta \approx 1.1$). This insight is clearly provided by the test information function. Items 5 and 6 are useful for separating examinees at high-ability levels. However, the item and ability parameters of scientific literacy reported in this study only pertain to the test responses by high school student that study physics in static fluid concept. Overall, the approach and findings of the current study may be used to develop and improve testing and enhance its sensitivity and effectiveness within a given range of abilities. Test developers can analyse item and ability parameters using IRT. Moreover, using IRT with the item parameters held constant, the same group of students can be tested before and after instruction to determine the learning gains in ability. The current results and approach can directly benefit anyone who uses the assessment, to gain improved accuracy of diagnosis. Other than that number of student participant should be added more to increase better precision of assessment test.
6. References

[1] Maulana A and Suhandi A 2017 AIP Conf. Proc. 1848(050004)
[2] Hambleton K, Swaminathan H and Rogers H 1991 Fundamentals of Item Response Theory (Newbury Park, CA: Sage Publications Inc.)
[3] Hambleton R K and Jones R W 1993 Educ. Meas. 12(38)
[4] MacDonald P and Paunonen S 2002 Educ. Psychol. Meas. 62(921)
[5] Kolen M J and Brennan R L 2004 Test Equating, Scaling, and Linking Methods and Practices, 2nd ed. (New York, NY: Springer Science and Business Media)
[6] Valois P, Houssemand C, Germain S and Abdous B, 2011 Proc. Soc. Behav. Sci. 15(552)
[7] Harris D 1989 Educ. Meas. 8(35)
[8] Organisation for Economic Co-Operation and Development (OECD) 2013 PISA 2015 Draft Science Framework (Paris: OECD)