Analysis of Quality of Critical Thinking Skills Test Based on Item Response Theory Using R-Program

Kaharuddin Arafah[1*], Burhanuddin Arafah[2,]AzhariahNur B. Arafah[3], Kasmawaru Kasmawaru[4]
[1]Physics Department at Faculty of Natural Sciences, State University of Makassar, Indonesia.
[2]English Department at Faculty of Cultural Sciences, Hasanuddin University of Makassar, Indonesia
[3]Psychology Department, Faculty of Medicine, Hasanuddin University of Makassar, Indonesia
[4]Department of Informatics Engineering, STMIK Dipanegara of Makassar, Indonesia
*kahar.arafah@unm.ac.id

ABSTRACT

This research aims to describe the quality of critical thinking skill test of fluid mechanics material in senior high schools in Makassar City. The emphasis is the content validity aspect, the characteristics of each item for the One-Parameter Logistic (1PL) model, the Two-Parameter Logistic (2PL) model, dan the Three-Parameter Logistic (3PL) model. It is a descriptive quantitative research with the subject of all student responses to the critical thinking skills test of fluid mechanics material at senior high schools in Makassar City. There were 726 student responses. The data were collected online by google form and analyzed by using descriptive quantitative technique. The results showed that the critical thinking skill test of fluid mechanics material had fulfilled the content validity. Analysis of the quality of CTS test items showed that the models of 1PL, 2PL, and 3PL were all consistent, showing that the test items were mostly able to discriminate the high abilitytestees from the low abilitytestees. Of the three logistic model approaches used to estimate the parameters of item difficulty level, item discrimination, and guess factor, the 3PL model is better than the other twomodels.

Keywords
1PL model; 2PL model; 3PL model; Critical thinking skills; Item response theory; R-Program; Quality; Test

Introduction

The 21st century learning including physics learning was characterized by 4C. Teachers must understand the 4C: creativity and innovation, collaboration, communication, and critical thinking and problem solving (Arafah, Rusyadi, Arafah&Arafah, 2020). The indicators of Critical thinking skill (CTS) are interpretation, analysis, evaluation, conclusions, explanations, and self-regulation (Facione, 2013). In addition, teachers should have digital literacy awareness consisting of information literacy, media literacy, and ICT literacy as their standard reference in providing instruments and making assessments, both in the affective, cognitive, and psychomorphic realms (Andi &Arafah, 2017; Arafah, Arafah&Arafah, 2020). To meet the demands of 21st century learning, a critical thinking skill instrument was developed in fluid mechanics material in senior high school (Arafah&Kaharuddin, 2019).No researcher has ever analyzed the instrument on fluid mechanics material. Even if there is, it is only limited to the classical theoretical approach in determining the discrimination power and difficulty level of items (Arafah&Setiyawati, 2020). This research was conducted to explore items characteristics by Item Renponse theory (IRT) approach. In this case, the One-Parameter Logistic (1PL) model, the Two-Parameter Logistic (2PL) model, and the Three-Parameter Logistic (3PL) model are used. This research aims to reveal the parameters of difficulty level of item (b_i), discrimination of item (a_i), and testee guess answers (c_i). This research also reveals the information function of item and the information function of test. Apart from these parameters, the item characteristic curve (ICC) estimation is also important in thisresearch.

Literature Review

The 1PL model is an estimation model of test item parameter reviewing difficulty level of item (b_i) by assumption that the discrimination of (a_i) is the same for all items and the guess answer (c_i) is zero. Preparatory to the analysis is to carry out the requirement test on the test item instrument. This requirement test is in the form of a unidimensional test and a local independence test (Naga, 2010). The unidimensional assumption is fulfilled if the items in the test instrument only measure one ability of the testees. Furthermore, local independence test is carried out to determine whether or not the testee responses to different test items are independent. The assumption of local independence depicts that the responses between one item and another are not related. The responses to an item from one testee with other testees are also not interconnected (independent). The equation for the 1PL model is described by equation 1 below (Umobong,2017).

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

Where D = 1,7

θ = Ability of testee

b_i = difficulty level of item

The 2PL model is a model that focuses on both the difficulty level (*bi*) and discrimination power (*ai*). For this reason, the guess factor is still assumed to be zero or there is no guess. It means that the chance of testees to answer a test item correctly is determined by the two characteristics of the item: the difficulty level of the test item and the test item discrimination. The equation for the 2PL logistic model is written as equation 2 (Naga,2010).

$$P_i(\theta) = \frac{e^{Da(\theta-b_i)}}{1+e^{Da(\theta-b_i)}} \quad (2)$$

Where D = 1,7

θ = Ability of testee

b_i = difficulty level of item

a = itemdiscrimination

The 3PL model is determined by three item parameters: difficulty level (*bi*), discrimination power (ai) and guess answer (*ci*) which are controlled jointly. It means that the odds of testees to answer a test item correctly is determined by the three characteristics of the item. As interpretation in 3PL model it also applies that the *bi* value or item difficulty index ranges from -2 to 2 and the $a_i$value ranges from 0 to 2. For the guess answer parameter as an odds measure for the testees to guess correctly, a test is expected to have little odds of guessing even close to zero. The equation for the 3PL logistic model is written as follows (Naga,2010).

$$P_i(\theta) = \frac{c_i + e^{Da(\theta-b_i)}}{1+e^{Da(\theta-b_i)}} \quad (3)$$

Where D = 1,7

θ = Ability of testee

b_i = difficulty level of item

a = itemdiscrimination

c = guess factor

The test quality aspect estimated in this research is the test information function. The test information function shows how much information a test instrument provides if it is given to the testees of certain ability.

Methodology

This research is a descriptive study with quantitative approach. The subjects were all answer sheet data of State Senior High School (SMAN) students in Makassar City who had responded to the given CTS questions in 2020. There were 726 answer sheets. The data analysis technique was theoretical and empirical data analysis. The theoretical analysis was to examine the test items with Expert Judgment (Arafah, Thayyib, Kaharuddin, & Sahib. 2020). For this purpose, the content validity is calculated using Gregory technique(2015).

Contentvalidity = $\frac{D}{A+B\ C+D}$, with;

		EXPERT JUDGE #1	
		Weak Relevance (item rated 1 or 2)	Strong Relevance (item rated 3 or 4)
EXPERT JUDGE #2	Weak Relevance (itemrated1 or 2)	A	B
	Strong Relevance (item rated 3 or 4)	C	D

Figure 1. Interrater agreement model for content validity

Content validity calculation of the CTS test on Fluid Mechanics material using Gregory equation results in an internal consistency coefficient between the two experts of 0.82. This figure is in the high category. Thus the CTS test can be used to collect data about critical thinking skills in senior high school students in Makassar City. The empirical analysis was to examine the characteristics of each test item based on Item Response Theory. This analysis was carried out using R i386 Vr 3.5.1 software program developed by Ihaka& Gentleman (1996).The one-parameter logistic model is a parameter estimation model of test items reviewing the difficulty level of item (b_i) assuming that the discrimination of (a_i) is the same for all items and the guess answer (c_i) is equal to zero. The local independent test, unidimensional, and the model fit test were all fulfilled. In the 2PL model, the difficulty level (b_i) and discrimination of (a_i) were calculated by assuming that the guess factor was equal to zero, while in the 3PL model the difficulty level (b_i), the discrimination of (a_i), and the guess answer (c_i) were controlled concurrently. Both the 1PL, 2PL and 3PL models display the iteminformation function and the test information function.

Results

The aspect of test quality estimated in this research is the test information function to show the extent the information given by a test instrument when it is used to test the testees with certain ability. The test informationfunction anditeminformationfor1PLmodelcanbeseen in Figure 2.

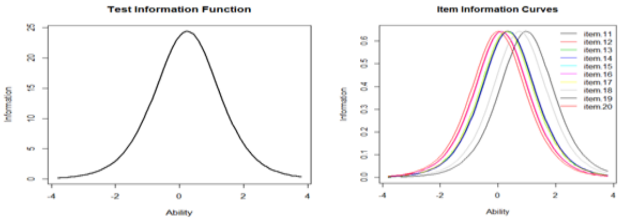

Figure 2: Curve of test information function and item information based on 1PL model

The test information function curve shows that if the estimation is made based on 1PL model, the CTS test of Fluid Mechanics material provides maximum information when applied to testees whose abilityθ between (-1) - (+1). The information function of item I (θ) is a description of test information function. Basically, the test information function is an accumulation of the item information function. The largest item information function is shown by item 14 with I (θ) = 0.69 and θ = (+0.46) - (+0.489).

Furthermore, if the estimation based on 2PL model, the CTS test of Fluid Mechanics material provides maximum information if it is applied to testees whose ability between (-1) - (+1). The test information function and item information based on the 2PL model are shown in Figure 3.

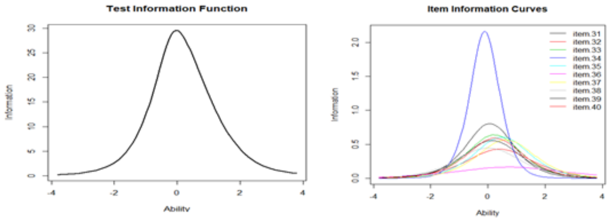

Figure 3. Curve of test information function and item information based on 2PL model

The test information function curve shows that if CTS test of Fluid Mechanics material isestimated based on 2PL model, it will provide maximum information when imposed on testees with ability between (-1) - (+1). Based on the curve in Figure 3, the largest item information function is indicated by item 34 with I (θ) = 2.33 and θ = (-0.079) -(0). Below is presented the estimation results of the characteristic of information function and item information function curve of the 40 items of CTS test of fluid mechanics in senior high schools (figure 4). Based on the curve in Figure 4, the maximum information is shown by the testees whose abilityθ between (+1) - (+1.178) with information value of test I (θ) = 60.05.

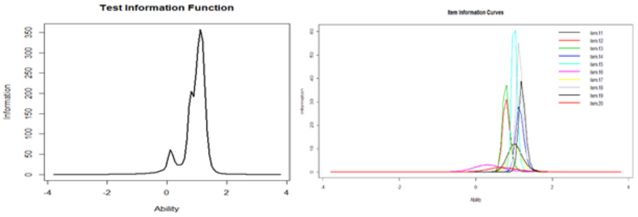

Figure 4. Curve of Test information function and item information based on 3PL model

This information is shown by item 15. This finding is in line with Mardapi (2017) that the information function is useful if the test items fit themodel.

Before analyzing the item characteristics, fitness test is conducted between the 1PL, 2PL and 3PL models. For this purpose, Akaike's information criterion (AIC) value and Bayesian information criterion (BIC) value are needed. The AIC and BIC values of each parameter model are presented in table1.

Table 1. Likelihood ratio for CTS test based on the 1PL, 2PL, and 3PL models

Models	AIC	BIC	log. Lik	LRT	df	p.value
1PL	30162.49	30350.64	-15040.25			
2PL	29986.27	30353.38	-14913.13	254.22	39	< 0.001
2PL	29986.27	30353.38	-14913.13			
3PL	28673.92	29224.59	-14216.96	1392.34	40	< 0.001
1PL	30162.49	30350.64	-15040.25			
3PL	28673.92	29224.59	-14216.96	1646.57	79	< 0.001

Source : Data analysis of program R i386 Vr.3.5.1

Table 1 shows that the parameter model which shows the lowest BIC value is 3PL (29224.59) compared to 2PL model (30353.38) and 1PL (30350.64) model. This shows that the most appropriate model used in this research is the 3-parameter logistic (3PL) model. Based on the prerequisite test,_the CTS test instruments of Fluid Mechanics in senior high school have fulfilled the aspects of unidimension, local independence and model fit test. Furthermore, from the 40 items of CTS test that had been tested, 19 items (47.5%) were in medium category with difficulty level value between (-0.840) - (+ 0.701), 13 items (32.5%) in difficult category with difficulty level value between +1.018 to +1.753), and 8 items (20.0%) in easy category with difficulty level value between -3.492 to -0.839. The discrimination parameter is assumed to be the same for all items, 0.45 and the guess answer is equal to zero for 1PL model. This finding is in line with Baker (2001) that the difficulty level parameter of item (*bi*) on ability to respond correctly to an item is 0.5. The greater the value parameter of *bi*, the greater the ability needed to answer the item correctly. The item characteristic curve (ICC) stating the relationship between the odds of testees to answer *Pi* (θ) correctly and the ability test (θ) on 1PL model is shown in Figure5.

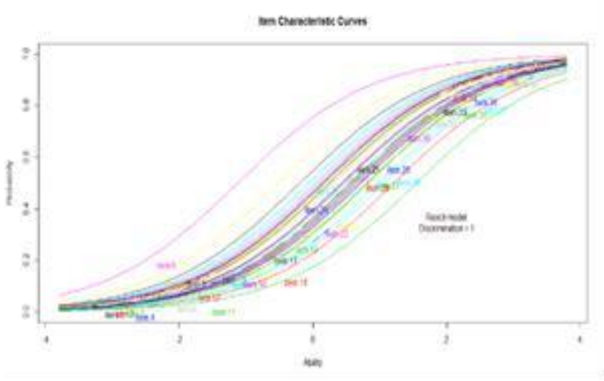

Figure 5. Characteristic curve of 40 test items in 1PL model

The item characteristic curve in Figure 5 above shows that the curve of the 40 test items moves from left to right, increasing continuously. The lowest asymptote approaches the normal ogive but never goes to zero and the highest asymptote approaches one. The highest curve is the curve for item 6 which describes that this item has odds of being answered correctly by moderate ability testees.Furthermore,thelowestcurveistheitem 25 which describes items categorized as very difficult. This Item 25 has odds of being answered correctly only by high-ability testees. Estimation results of the CTS test item discrimination for 2PL model indicates that 23 test items or 57.5% have low value discrimination with an index ranging from (+0.842) - (+1.656). 14 test items or 35.0% have moderate value discrimination with a discrimination index ranging from (1.705) - (+2,469). 3 test items or 7.5% have high value with an index ranging from (+3.101) - (+3.404). This shows that there are only 3 (7.5%) test items which could discriminate the high-ability testees and the low-ability testees. However, there are still 23 test items capable to discriminate well between the high-ability testees and the low-ability testees. As to item difficulty level is found 14 test items with easy category level of difficulty with the index between (+0.044) - (+0.072). 25 test items or 62.5% have moderate category level of item difficulty with the index ranging from (+0.073) - (+0.101). Only 1 test item or 2.5% has difficult category level with an index (+0.131). Item characteristic curve (ICC) denotes the odds thetesteeswithcertainabilityhavetoanswerthe items correctly. Figure 6 below show the item characteristic curve of 2PL model.

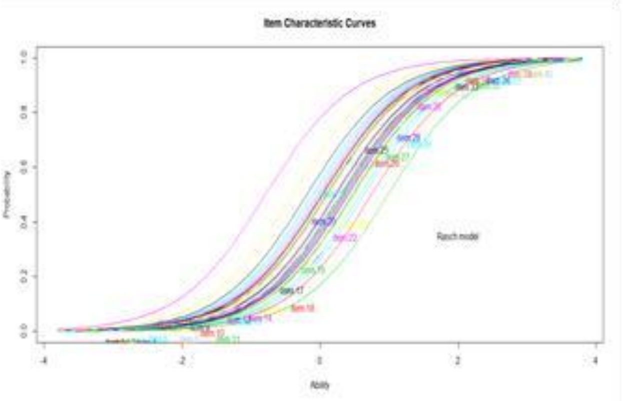

Figure 6. Characteristic Curve of 45 test items with the 2PL Logistic Model Estimation

The curve model expected to have ideal items is the sloping letter S. Almost all of the items are in the form of sloping S. Thus, the parameter estimation of 2PL logistic model shows that the greater the ability of testees, the greater the odds to answer the test items correctly. The estimation results of discrimination parameter in 3PL model shows that 19 test items have low discrimination power with discrimination values ranging from (+1,901) - (+6,659). 19 test items have moderate discrimination power with discrimination index ranging from (+6,660) - (+11,418). There are only 6 test items that have high discrimination power with index ranging from (+11,420) - (+16,177). This shows that 21 test items or 52.5% are able to discriminate the high-ability and low-ability testees. Meanwhile, 19 test items or the remaining 47.5% are unable to discriminate between high and low ability testees. Likewise the analysis results of 40 test items in which 3 test items or 7.5% are found in easy category with difficulty level value between (+0.185) - (+0.399). 10 test items or 25.0% are in moderate category with difficulty index between (+0.340) - (+0.984), and 27 items are in difficult category. The results of the guess answer analysis show that of the 40 CTS test items, there are 28 test items or 70.0% with the guess answers functioning effectively, with the guess answer value from (+0.135) - (+0.269). The remaining 12 test items or 30.0% showthe guess answers do not work. The following shows the characteristic curves of 3PL model test items.

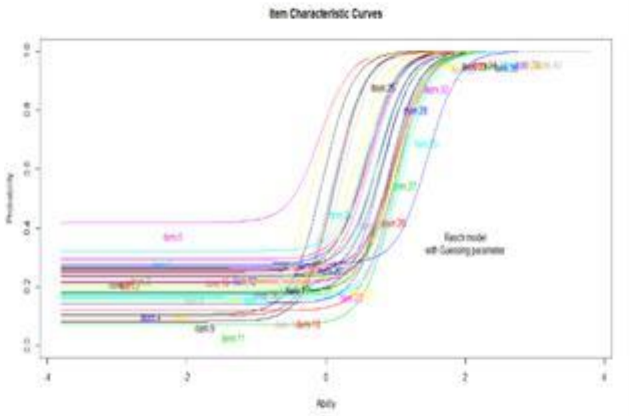

Figure 7. Characteristic curve of 40 test items based on 3PL model estimation

The characteristic curve of the 40 test items of CTS shows the shape of quite sloping S in general. This shows that the test items have discrimination power, difficulty level, and guess odds working well. These fairly ideal items are represented by items 26, 28, and 30. The greater the testee ability, the greater the odds to answer the test item correctly and the odds to guess answers is zero. Furthermore, based on the estimation of 3PL model, 30% of test items still denote high enough odds to guess. The test items should becorrected.

Discussions

The item information function generated by 1PL model is different from 2PL model and 3PL model. The information conveyed by the 3PL model is more compared to the 2PL model and the 1PL model. It indicates that each item provides different information. For example, if the CTS test of Fluid Mechanics material is estimated based on the 1PL model, then the largest item information function is indicated by item 14 with I (θ) = 0.69 and (θ) = (+0.46) — (+0.489). If the estimation is based on the 2PL model, the maximum information will be given to testees with ability (θ) between (-1) - (+1). If the estimation is based on the 3PL model, then the maximum information is indicated by testees with ability (θ)between (+1) - (+1.178) with test information value of I (θ) 60.05. Hence, the test information function is theaccumulation of item information function.

The greater the parameter value of item difficulty level (b_i), the greater the ability required to answer the item correctly. *Item characteristic curve* (ICC) states the relationship of the odds of testee to answer correctly Pi(θ) with the ability of testee (θ). Based on figure 5, the curve for item 6 has odds to be answered correctly by testee with moderate ability. Meanwhile the item 25 has odds to be answered correctly only by testee with high ability, because the item is verydifficult.

The characteristic curve of the 40 items of CTS test shows that generally the S-shaped curve is quite sloping. This indicates that test items with discrimination power, difficulty levels, and guessing odds function properly. The greater the ability of testee, the odds to answer correctly the test items is also greater and the odds of guessing answer is equal to zero.

Conclusion

Based on the research results and discussion, the conclusions are drawn as follows;

The CTS test for Fluid Mechanics material can be used to measure the critical thinking skills of senior high school students in Makassar City because it has fulfilled the Content Validity. Analysis of the CTS test items quality shows that both the 1PL, 2PL, and 3PL models are consistent, showing that generally the test items are able to discriminate the high and low-ability testees. Of the three logistic model approaches used to estimate the parameters of item difficulty level, item discrimination power, and guess factor, the 3PL model is the best among other models. Thus it can be said that the CTS test for Fluid Mechanics material has goodquality.

Limitations and Future Studies

This study only investigated three indicators of process skills presented by Facione (2013): interpretation, analysis, and evaluation. Other indicators were not investigated. Researchers interested in continuing this research are suggested to examine the overall indicators of critical thinkingskills.

Acknowledgement

This research was conducted with the support of many people, particularly the principals in Makassar City. Therefore, great thanks go to the head of SMAN 2 Makassar, SMAN 4 Makassar, SMAN 8 Makassar, SMAN 9 Makassar, and SMAN 22 Makassar for their assistance in data collectionprocess.

References

[1] Andi, K., &Arafah, B. (2017). Using needs analysis to develop English teaching materials in initial speaking skills for Indonesian college students of English. *The Turkish Online Journal of Design, Art and Communication (TOJDAC), Special Edition*, 419-436.

[2] Arafah, B., &Kaharuddin. (2019). The Representation of Complaints in English and Indonesian Discourses. *Opción, 35*, 501-517.

[3] Arafah, K., Arafah, A. N. B., &Arafah, B. (2020). Self-Concept and Self-Efficacy's Role in Achievement Motivation and Physics Learning Outcomes. *Opción*, 36, (27),1607-1623.

[4] Arafah, K., Rusyadi, R., Arafah, B., &Arafah, A. N. B. (2020). The Effect of Guided Inquiry Model and Learning Motivation on the Understanding of Physics Concepts. *Journal of Talent Development and Excellence*, *12*(1),4271-4283.

[5] Arafah, B., Thayyib, M., Kaharuddin, & Sahib, H. (2020). An anthropological linguistic study on Maccera'Bulung ritual, *Opción*, 36, (27), 1592-1606.

[6] Arafah, A. N. B., &Setiyawati, D. (2020). Volunteerism in Sub-District Social Welfare Worker in Dosaraso Halfway House. *International Journal ofPsychosocial Rehabilitation*, *24*(Special Issue 1), 357–362. doi: 10.37200/ijpr/v24sp1/pr201165.

[7] Mardapi, D. (2017). *PengukuranPenilaiandanEvaluasiPendidikan*. Yogyakarta: ParamaPublishing.

[8] Hambleton, R.K., Swaminathan, H & Rogers, H.J. (1991). *Fundamental of item response theory*. Newbury Park, CA: Sage PublicationInc.

[9] Ihaka, R., & Gentleman, R. (1996). A Languge for Data Analysis and Graphics, *Journal of Computational andGraphical statistics*, Volume 5 Number 3 : 299-314.

[10] Gregory, R. J. (2015). *Psychological Testing: History, Principles, and Applications*,Global Edition,England: Pearson EducationLimited.

[11] Baker, F. B. (2001). *The Basics of Item Response Theory*, United States of America: ERIC Clearinghouse on Assessment and Evaluation.

[12] Facione, P. A. (2013). *Critical Thinking: What It Is and Why It Counts*,1–28.

[13] Umobong, M. E, (2017). The One-Parameter Logistic Model (1PLM) and Its Application in Test Development. *Advances in Social Sciences Research Journal*, (424)126-137.

[14] Naga, D. S. (2010). *TeoriSekorPadaPengukuran Mental*, Jakarta Barat: PT NagaraniCitrayasa.