Exploring the final examination test item characteristics of Pancasila and civic education

Syarief Fajaruddin a,*, Heri Retnawati b, Eri Yusron c,d, Vinni Sofyaningsih e,d

a Universitas Negeri Yogyakarta, Indonesia 
b Universitas Pendidikan Indonesia, Indonesia
c Madrasah Tsanawiyah Negeri 3 Gunungkidul, Indonesia

* syarief@uny.ac.id; heri_retnawati@uny.ac.id; eyusron98@gmail.com; vinni.soffi@gmail.com

1. Introduction

In the era of globalization with the advancement of modern technology, education becomes the top priority in order that the development of human resources quality can be the most significant aspect [1]. The reason is that human resources quality depends on education quality [2]–[4]. Well-qualified human resources are the main asset within the development of a nation [5]. In addition, the development of a nation can be viewed from how far the existing education has advanced itself [6]–[8]. In relation to the statement, one of the determiners within the success of education is the high influence by the teacher capacity in implementing the learning activities [9], [10]. Therefore, learning activities are expected to be effective [11], interesting [9], [12], and fun [13]. In addition, another determiner is the necessity for developing various learning models to improve the learning quality [14] and the students’ learning results [14]. To identify how far the learning quality and the learning results of the students have progressed, the teachers should conduct an evaluation.

Learning evaluation is one of the ways for attaining information with regards to the overall gain of the students in the aspects of knowledge, concept, attitude, value, and even process skills [15]. Through evaluation, teachers will be able to identify and understand both the individual and the communal achievements of the students [16]. Unfortunately, in Indonesia, there are still a number of issues that have disrupted the learning evaluation. According to Rotama et al. (2020), many teachers still put forward the cognitive aspects of the students and the instruments that teachers design have not undergone any validation process. In addition, the test items are rarely reviewed in terms of
validity and reliability, material, construction, language, and even test item analysis after the test items have been administered based on the difficulty level, the item discrimination capacity, and the dummy analysis [17]. This issue appears to the surface due to the limited staff [18] and the busy teaching schedule, leading to the insufficient time for performing the test item analysis and also the lack of knowledge and understanding on the part of the teachers with regards to the test item analysis that should be conducted [17], [19].

The Republic of Indonesia Government Regulation No. 19 of 2005 on National Education Standard mentions that educational assessment on the elementary and high school degree consists of learning results assessment by the educators, learning results assessment by the educational units and learning results assessment by the government [20]. On the contrary, article 64 verse 1 of the same regulation states that the learning results assessment by the teacher as having been intended by article 63 verse 1 is conducted continuously to monitor the process, the progress, and the improvement of the learning results in the form of daily test, mid-semester test, final examination test, and class promotion test. The process of evaluating the students’ learning results can be administered by performing the test technique and the non-test technique [18], [21], [22]. Most of the time, teachers implement the test technique in the form of daily tests, mid-semester tests, and final semester tests [18]. All these tests can be either subjective tests or objective tests. In general, the subjective test takes the form of an essay whole the objective test takes the form of a true-false, multiple-choice test, matching test, and completion test [23].

A test can be considered good if the test meets the criteria of validity, reliability, objectivity, practicability, and economics [23], [24]. Thus, the test is expected to provide accurate measurement results. In the context of the study, the test that has been administered is the multiple-choice objective test. According to [25], this type of test offers a number of benefits, namely: (a) being able to measure the learning results objectively; (b) providing faster correction rate; (c) providing faster notification on the scores of the students, and (d) being able to be turned into the test item bank. As an alternative, according to Mania et al. (2020); Slamet dan Maarif (2014), multiple-choice test items offer the following benefit: (a) the test items are easier to analyze; (b) the test items cover many learning materials; (c) all indicators can be met; and (d) students’ capacity can be measured in accordance with the desired domain and the difficulty level [26], [27]. Unfortunately, despite those benefits, the multiple-choice test items still suffer from several weaknesses, namely: (a) the designing time is quite demanding; and (b) the designing process takes huge fund resources [25]. Not to mention, the multiple-choice test items are less able to describe the process since they only measure cognitive skills. Therefore, through the multiple-choice test items, the students are able to answer the test items without having to analyze them, and thus, the capacity of the students cannot be completely described [26].

2. Method

The study is descriptive quantitative research that pursues the document analysis to view the characteristics of the Even Semester Examination test item characteristics of Pancasila and Civic Education for Grade VII of MTs Negeri 3 in the Regency of Gunungkidul, the Province of Yogyakarta Special Region, for the 2020/2021 Academic Year (hereinafter shall be referred to MTs Negeri 3 Gunungkidul). The approach adopted within the study is both the qualitative and the quantitative approach, and both approaches have been adopted to view the quality of the final semester examination test item [28]. The subjects within the study are all students from Grade VII of MTs Negeri 3 Gunungkidul with a total number of 161 people. Then, the objects of the study are the responses from the even semester examination test item of Pancasila and Civic Education for Grade VII in the 2020/2021 Academic Year with a total number of 40 items. For the test scoring, the researchers have administered the polytomous scoring with the ordinal scale of 1-2-3-4.

Furthermore, the data-gathering technique that has been implemented is the documentation technique. The documentation technique is implemented to attain the data in the form of even semester examination test item of Pancasila and Civic Education for Grade VII, the answer keys of the test items, and the answer sheets of all Grade VII students from the given subject. The test item analysis using the IRT should meet the assumptions that have been required, and these assumptions are unidimensional assumption, local independent assumption, and invariant parameter assumption. The unidimensional assumption asserts that each test only measures one skill. Thus, the statement implies that every test item only measures one skill of the test takers [29]. In other words, the
probability of an item response serves as the single latent characteristic of the test takers [30]. Therefore, a test that has been administered is expected to measure one character or one skill. Then, to meet the unidimensional assumption, the factor that has the most dominant influence on the test performance should be compared to the objective of the test design. If the dominant factor that appears to the surface already meets the objective of the test design, then the unidimensional assumption has already been met. Within the context of the study, the unidimensional assumption testing is conducted by using the SPSS Program.

Next, the local independent assumption defines that the performance of an individual over a test item does not influence the performance of the individual on another test item. This assumption will be met if the response of the test takers on a test item does not influence the response of the test takers on another test item [31]. Last but not least, the invariant parameter assumption defines that the test item characteristics do not depend on the skill parameter distribution of the test takers, and the parameter that becomes the characteristics of the test takers does not depend on the test item characteristics [31]. The implication of this assumption is that the skills of the test takers will not change only because they respond to the test items with different difficult levels [29].

The analysis of the test item is conducted by using the R Program. With regards to the statement, the criteria of item quality within the study refer to the requirements that have been outlined by Hulin et al. (1983), which consists of “Good,” “Poor,” and “Very Poor.” These criteria can be broken down into specific explanations as follows: (a) an item will belong to the “Good” category if the item fits into the model, its difficulty index ranges between -2.0 and 2.0, and its item discriminative index ranges between 0.0 and 2.0; (b) an item will belong to the “Poor” category if the item less fits into the model, its difficulty index ranges is < -2.0 or > 2.0, and its item discriminative capacity is >2.0; and (c) an item will belong to the “Very Poor” category if the item does not completely fit into the model.

3. Results and Discussion

The results of the study show several information such as item response theory assumption test results, fitness model, item parameter coefficient, item characteristic curve plot, and theta respondent. Within the assumption test, the unidimensional procedure should be conducted first. The unidimensional assumption test is conducted through the exploratory factor analysis using the SPSS program. One of the aspects that should be given attention in performing the exploratory factor analysis is the fulfillment of the sample sufficiency. To identify the sample sufficiency, the values of Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO MSA) can be consulted in Table 1. Based on the analysis results, the KMO value of the instrument is 0.801 with $p < 0.05$. This value is higher than the KMO reference value that has been required, namely $\geq 0.50$. In other words, the sample size, 161 respondents, within the analysis has been sufficient. The unidimensional test results are available in Table 1.

The results of factor analysis toward the instrument displayed in Fig 1 show that the instrument within the study only has one dimension. The unidimensional characteristic is apparent since there is only one factor whose eigenvalue has been higher than 1. The eigenvalue of the first factor is 0.871, while the eigenvalue of the second factor is lower than 1.000. In the meantime, the remaining eigenvalues are lower than 1.000. Furthermore, within the IRT assumption test, the local independence procedure is also administered. The local independence criteria will be met if the correlation values on each item are lower than 0.200. The local independence test itself is conducted using Yen’s Q3. The results of the local independence test show that the highest correlation value is 0.200, while the correlation value of the remaining item is lower than 0.200. Therefore, this finding suggests that the IRT assumption test criteria have been met.

<table>
<thead>
<tr>
<th>Table 1. KMO and Bartlett’s test</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>KMO and Bartlett’s Test</strong></td>
</tr>
<tr>
<td>Kaiser-Meyer-Olkin Measure of Sampling Adequacy</td>
</tr>
<tr>
<td>Bartlett’s Test of Sphericity</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Syarief Fajaruddin et al. (Exploring the final examination test item characteristics ...
Next, the results of the study also display the model fitness. In viewing the model fitness for the analysis, the researchers compare the Akaike Information Criterion (AIC) value, the Bayesian Information Criterion (BYC) value, and the log Lik value. The lower these values are, the more fit the model in analyzing the data that will be used. The results of the comparison on the model fitness are available in Table 2 and Table 3. Table 2 displays the information on comparing the Rasch Model and the 1PL (1-parameter logistic) Model.

![Fig. 1. Unidimensional Test Results](image)

The results of the comparison show that significantly (p.value < 0.001) the AIC value, the BIC value, and the log.Lik values within the 2PL (2-parameter logistic) are lower than the values in the Rasch Model. These findings show that the model fitness comparison between the Rasch Model and the 2PL Model for the data analysis tends to favour the use of the 2 PL Model. Consequently, the researchers should compare the fitness between the 2PL Model and the 3PL (3-parameter logistic) Model. The results of this comparison are available in Table 4.

### Table 2. Likelihood ratio table

<table>
<thead>
<tr>
<th></th>
<th>AIC</th>
<th>BIC</th>
<th>log.lik</th>
<th>LRT</th>
<th>df</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rasch Model</td>
<td>5362.13</td>
<td>5488.46</td>
<td>-2640.06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2PL Model</td>
<td>5144.82</td>
<td>5391.33</td>
<td>-2492.41</td>
<td>295.31</td>
<td>39</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

### Table 3. Likelihood ratio table

<table>
<thead>
<tr>
<th></th>
<th>AIC</th>
<th>BIC</th>
<th>log.lik</th>
<th>LRT</th>
<th>df</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 2PL</td>
<td>5144.82</td>
<td>5391.33</td>
<td>-2492.41</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 3PL</td>
<td>5149.24</td>
<td>5519.01</td>
<td>-2454.62</td>
<td>75.58</td>
<td>40</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Next, Table 3 shows that the AIC value, the BIC value, and the log.Lik values on the 2PL Model are significantly lower than the values in the 3PL model. Thus, this finding explains that the data within the study are more appropriate to be analyzed using the 2PL Model. Furthermore, Table 4 informs about the difficulty parameter coefficient and the item discriminatory capacity using the 2PL Model a good difficulty index range between -2.00 and +2.00. From the results of the 2PL model analysis toward the test item of Pancasila and civic education final semester examination, it is found that the most difficult item is item number 18, which difficulty coefficient has been 3.422, while the easiest item is item number 27, which difficulty coefficient has been 3.076. The difficulty degree of the 40 test items itself ranges between -4.00 and +4.00. In general, the item difficulty falls into the coefficient -0.979 with the difficulty standard deviation of 1.139.
This finding shows that overall, the instrument has a good difficulty level in each item. In relation to the statement, the 2PL logistic model analysis yields different discriminatory capacities for each item. In this regard, the results in Table 4 show the discriminatory capacity range between 0.0079 and 4.891. The item with the lowest discriminatory capacity is item number 18, which index is 0.079, while the item with the highest discriminatory capacity is item number 30, which index is 4.891. There are several ICCs (Item Characteristic Curve) from several items, and these curves are available in Fig 2.

Fig 2 refers to the ICC for the item number and item number 30. The ICC for item number 10 describes that the item has a high difficulty index, as is apparent from the very slopy curve line.
Therefore, it is apparent that the probability for the respondents with theta 2 to respond to the test item correctly is only 0.40. Thus, the test item is moderate in discriminating the respondent capacity. In the meantime, the ICC for test item number 30 shows that the curve line falls between -2.00 and -1.00, which is very steep. This finding shows that the item discriminatory capacity for the test item number 30 is poor and, thus, it has a low difficulty index.

The item analysis using the IRT should meet the three assumptions that have been required [30]. The assumptions that have been generally used in the IRT models are the unidimensional assumption, the local independent assumption, and the invariant parameter assumption [31], [33], [34]. The unidimensional test is conducted in order to identify whether a test measures only one trait or not [30]. In this regard, the results of the analysis show that the instrument has been confirmed to have only one dimension, namely measuring the student capacity in Pancasila and Civic Education. Then, in the local independent test, the results of the analysis show that the highest correlation value is 0.200 while the remaining correlation value is lower than 0.200. Thus, this finding implies that the IRT assumption test within the instrument has already been met. This finding is in accordance with the results of a study by Hambleton dan Swaminathan (1985), which state that if the covariant value of the students’ skills group is closer to zero, then the local independent assumption test criteria have already been met. In addition, the local independent assumption test will be met if the test has been confirmed to be unidimensional [35]–[37]. The local independent test is performed to identify the students’ response on a test item and the response should not be dependent on their response to the other item.

The fitness model that has been adopted in the study is the 2PL Model. This finding is attained after the instrument is compared in terms of compatibility among the 1PL (Rasch) Model, the 2PL Model, and the 3PL Model. The lower the values in the data analysis are, the fitter, the model, will be in analyzing the data that will be used within the study. The statement is in line with the results of a study by Jafar, which state that the parameter model that shows the lowest Akaike Information Criterion value is a fit for use. The results of the test item analysis for the Final Semester Examination of Pancasila and Civic Education in Grade VII MTs Negeri 3 Gunungkidul using the 2PL Model inform about the difficulty level (bi) and the discriminatory capacity (ai). The results of the analysis show that the instrument, overall, has a good difficulty index for each item. These results have been confirmed with the 35 test items (87.50%) that belong to the “Good” category and 5 test items (12.50%) that belong to the “Poor” category. This conclusion is based on the range between -2.00 and +2.00 within the logit scale [34], [35], [39], [40]. In the meantime, the discriminatory capacity information within the instrument shows sufficient results. The statement is based on the argument by DeMars (2018), who states that the good discriminatory capacity range between 0.00 and +2.00. Therefore, it can be safely concluded that the test item instrument of Pancasila and Civic Education in MTs Negeri 3 Gunungkidul is able to differentiate between the high-performing students and the low-performing students.

<table>
<thead>
<tr>
<th>Table 5. Test Participant Capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>EAP</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>SD</td>
</tr>
<tr>
<td>Max.</td>
</tr>
<tr>
<td>Min.</td>
</tr>
</tbody>
</table>

4. Conclusions

The test item analysis for the Final Semester Examination of Pancasila and Civic Education in MTs Negeri 3 Gunungkidul shows that the test item is more appropriate to be analyzed by using the 2PL Model. Based on the analysis using the 2PL Model, it is found that the most difficult item is test item number 18 while the easiest test item is item number 27, with the degree of difficulty that falls into the range between -4.00 and +4.00. In general, the coefficient of the difficulty index is -0.979, with the difficulty standard deviation of 1.139. Thus, this finding implies that the instrument has a good difficulty index. In the meantime, the item discriminatory capacity falls into the range between 0.079 and 4.891. The item with the lowest discriminatory index is item number 18, while the item with the highest discriminatory index is item number 30.
Acknowledgment

The researchers would like to thank the Lecturer of Item Response Theory for the guidance and all the colleagues from PEP Doctoral Program Batch 2020 for the support. Then, the researchers would also like to thank Jean Antunes Rudolf Zico Ma’u for the willingness to translate the article.

Declarations

Author contribution: SF: Concepting and composing the manuscript; HR: Guiding and directing; EY: Processing the data; VS: Collecting the data

Funding statement: The author received no financial support for the research and/or authorship of this article

Conflict of interest: The author declares no conflict of interest.

Additional information: No additional information is available for this paper.

References


