REID (Research and Evaluation in Education), 9(2), 2023, 210-222
Available online at: http://journal.uny.ac.id/index.php/reid

Revealing the characteristics of Indonesian language test used in the national-standardized school examinations

Marwah Ulwatunnisa[1]*; Heri Retnawati[1]; Muhardis[2]; Eri Yusron[1,3]
[1]Universitas Negeri Yogyakarta, Indonesia
[2]Badan Riset dan Inovasi Nasional (BRIN), Indonesia
[3]Pusat Studi Pendidikan dan Kebijakan (PSPK), Indonesia
*Corresponding Author. E-mail: marwaulnisa@gmail.com

ARTICLE INFO

Article History
Submitted:
26 May 2020
Revised:
04 December 2023
Accepted:
20 December 2023

Keywords
discriminating power; Indonesian language test; item difficulty; school examinations; standardized test

Scan Me:

ABSTRACT

This study aimed to determine the characteristics of Indonesian language test used in the national standardized school examinations (*Ujian Sekolah Berstandar Nasional*, USBN). We used the response data of 218 students from a public senior high school in the Special Region of Yogyakarta, Indonesia, to the test consisting of two packages, A and B, in the 2018/2019 academic year to investigate the characteristics of the test and its items. Quantitative analysis using classical test theory (CTT) and one-parameter logistic item response theory (1-PL IRT) model was conducted to investigate the characteristics of the test and its items based on difficulty and discriminating power. The results of the study under the CTT showed that most test items in both package A and package B have difficulty in the easy category and the portion of items in the difficult category is no more than 10%. In addition, while the majority of test items in both test packages demonstrated good discriminating power, package B contained a high number of items with poor discriminating power (47.5%). Under 1-PL IRT, our study results indicated the dominance of items with difficulty level in the moderate category in both test packages. In addition to revealing the difficulty level of the test items, our study showed that the items in the difficult category under CTT and IRT were related to the topics of types of conjunction; spelling, grammar, and sentence structure; job application letter; and observation report text and its structure. The results of this study are expected to contribute to improving the quality of the Indonesian language test and the quality of learning on topics where students have difficulty based on test items in the difficult category.

To cite this article (in APA style):
Ulwatunnisa, M., Retnawati, H., Muhardis, M., & Yusron, E. (2023). Revealing the characteristics of Indonesian language test used in the national-standardized school examinations. *REID (Research and Evaluation in Education), 9*(2), 210–222. https://doi.org/10.21831/reid.v9i2.31999

INTRODUCTION

Learning objectives can be divided into three domains, namely the cognitive domain related to thinking competence, affective aspects related to attitude factors, and psychomotor related to psychomotor competence (Nurgiyantoro, 2016). To investigate the extent to which these learning objectives are achieved, assessment needs to be carried out. According to Gronlund et al. (2009), assessments are all kinds of procedures used to obtain information about student performance, which when associated with learning objectives means that it is to determine the extent to which students master the competencies as stated in the learning objectives. In other words, through assessment, we can ascertain how well students perform in their learning as reflected by their mastery of a number of competencies. Retnawati et al. (2017) suggest that assessment is important

ISSN: 2460-6995 (Online)

to determine educational success. The results of assessment in education have the main role of being useful in the process of further education.

Various assessments have been conducted at school, provincial, and national levels. One of these assessments is the national-standardized school examinations (or *Ujian Sekolah Berstandar Nasional*, USBN) – USBN is currently no longer held. USBN was carried out by educational units or schools and was focused on identifying the extent to which students have achieved a number of competencies that are expected to be mastered after they graduate from a certain level of education which is then used to provide recognition of the learning they have undertaken at that level of education (National Education Standards Board, 2018). Student achievement in USBN is not the sole determinant of student graduation from a level of education. However, given the standardized nature of the USBN, data of student achievement on the USBN can be used to support the derivation of education policies at local, regional and national levels; as a basis for school quality improvement; and as a basis for teachers to improve the quality of the learning practices they facilitate and the quality of their professional competence (Huber & Skedsmo, 2017). A number of subjects were tested on students in the USBN, one of which is Indonesian language, which is then referred to as the Indonesian language test. This test used the combination of selected response type questions or items in the form of multiple-choice and constructed response in the form of essay. The government through the National Education Standards Board (2018) regulated that the items on the test are a combination of 75 to 80 percent of test items constructed by subject teachers in education units or schools which were further consolidated with subject teacher forums and 20 to 25 percent of test items used as anchor items that have been constructed and provided by the government.

As part of a standardized test, the Indonesian language test used in USBN must certainly satisfy the criteria or standards of quality that have been set. It is widely acknowledged that some of the criteria or standards for the quality of a test or measurement instrument lie in the validity and reliability of the test (Ebel & Frisbie, 1991; Evers et al., 2013; Rafi et al., 2023). Given that the quality of a test or measurement instrument can also be influenced by the quality of the items that make it up, investigating the test items to determine the level of difficulty and item performance in discriminating the ability of test takers through item analysis is something that needs to be a major concern (Arruarte et al., 2021; Costello et al., 2018). Item analysis plays a strategic role in the development of a test or measurement instrument because the results obtained from item analysis can be used by test developers as a basis for reviewing and improving the test items developed through revealing the quality of the test items based on their characteristics (Moses, 2017). Two characteristics of test items that are usually taken into consideration in the development of tests or measurement instruments are the level of difficulty and discriminating power of an item. A test or measurement instrument is expected to be built by items that have a good power to distinguish the ability of test takers and have a certain level of difficulty that depends on the objective of the measurement. The investigation of the characteristics of test items based on these two aspects through item analysis is usually based on two measurement theories, namely classical test theory (CTT) and modern test theory based on test taker responses, also known as item response theory (IRT).

According to Mardapi (1999), classical test theory (CTT) is considered as a measurement theory that uses assumptions that are relatively simple and easy to understand. In analyzing test items through CTT, a good test should at least conform to three characteristics: item difficulty, item discrimination, and effectiveness of distractors (Maharani, 2020). In addition, Retnawati et al. (2011) proposed several item parameters related to CTT, they are the proportion of correct answers, difficulty level, reliability, discriminating power, and measurement errors. Items with high discrimination power are generally the most desirable, and the level of difficulty of the corresponding items is determined by the test objectives, including the distribution of anticipated abilities of the groups that are the test objectives (Hambleton et al., 1991). The scoring procedure in educational assessment in the CTT approach is based on the correct answer. If students correctly

answered the multiple-choice item, they were given a score of 1; if they answered incorrectly, they were given a score of 0. This kind of scoring procedure is stated with the total score obtained by students. This procedure does not consider the interaction between each student and the items (Retnawati et al., 2011).

Retnawati et al. (2011) explained that the difficulty of an item, which is symbolized by p, is one of the item parameters which is very useful for analyzing a test. By looking at the parameters of a test item, it will be known how good the test item quality is. The value of p of a multiple-choice item on a test is obtained by determining the result of the division between the number of test takers who answered the item correctly and the number of test takers. If the value of p of an item approaches 0, the item is perfectly difficult. If the value of p of an item approaches 1, the item is perfectly easy. When the p value of an item is equal to 0 or 1, the item is needed to be discarded as the item cannot distinguish student's ability. Allen and Yen (2002) suggested that in general the difficulty index of an item (p) should be between 0.3 and 0.7. At this interval, information about the ability of students will be obtained maximally. In designing a test or measurement instrument, it is necessary to consider the objective of the test or measurement instrument itself. In addition to item difficulty, as we mentioned earlier, item discrimination, which represents how well an item discriminates between students in terms of their ability based on their total test score (Moses, 2017; Rafi et al., 2023), is also a concern in CTT. One index that can be used to determine whether an item has good differentiating power is the point-biserial correlation coefficient (r_{pbis})(Moses, 2017).

In addition to being based on CTT, item analysis to identify the characteristics of test items can also be carried out under item response theory (IRT). Three basic assumptions in the IRT must be met. These assumptions are unidimensional, local independence, and the accuracy of the item characteristic curve (Hambleton et al., 1991). Retnawati (2014, pp. 12–13) explains that the IRT approach is an alternative approach that can be used in analyzing a test and its items. There are two principles used in this approach, namely the principle of relativity and the principle of probability. On the principle of relativity, the analysis unit of measurement is not students or items, but rather student performance relative to items. If β_n denotes n-th student's ability on the measured trait and δ_i is an index of the difficulty level of the i-th item associated with the measured ability, then it is not β_n or δ_i which is a unit of measurement, but rather the difference between student's ability relative to the level of difficulty of items ($\beta_n - \delta_i$). As an alternative, a comparison between the ability to the level of difficulty can be used. If the ability of students exceeds the item difficulty level, then the student response is expected to be correct, and if the student's ability is less than the item difficulty level, then the student response is expected to be wrong. In addition, in IRT, the principle of probability is of concern. Let us suppose that the n-th student's ability is denoted by θ_n and the difficulty level of the item is expressed by Δ_i. According to the principle of probability, if $\theta_n > \Delta_i$, the student is expected to answer the item i correctly, and if $\theta_n < \Delta_i$, the student is expected to answer the item i incorrectly.

Similar to item analysis through the CTT approach where we can identify item characteristics based on difficulty and differentiation, item analysis through the IRT approach also allows us to identify item characteristics based on these two aspects through several available parameter logistic models. The logistic parameter models in IRT even allow us to identify item characteristics beyond these two aspects, where we can identify item characteristics by considering the responses given by low-ability test takers (related to the pseudo-guessing parameter) and high-ability test takers (related to the carelessness parameter). Equation (1) represents the general equation of the parameter logistic models in the IRT approach; where $P(\theta_j)$ is the probability of the j-th test taker with ability level θ to answer a dichotomous item correctly, a_i is the discrimination parameter of item i, b_i is the difficulty parameter of item i, g_i is the maximum probability of a low-ability test taker to answer item i correctly (so-called the pseudo-guessing parameter), and u_i is the maximum probability of a high-ability test taker to answer item i correctly (so-called the carelessness parameter), and D is a scaling constant whose value can be set equal to 1 or 1.702 (Pardede et al., 2023, p. 92). The name of the models in Equation (1) depends on the number of parameters or characteristics involved.

When the model focuses on only one parameter, i.e., the difficulty parameter, it is called a one-parameter logistic (1-PL) model. This model is obtained by setting $g = 0$, $u = 1$, and a to the same value for all items.

$$P(\theta_j) = g_i + (u_i - g_i)\frac{\exp[Da_i(\theta_j - b_i)]}{1 + \exp[Da_i(\theta_j - b_i)]} \quad (1)$$

In the 1-PL model, the b value of an item (item difficulty) represents the level of ability required by test takers so that their probability of responding correctly to the item is 50%. For example, a parameter b of a test item of 0.3 indicates that the minimum ability level a test taker must have in order to have a probability of responding correctly to the item is 0.3. The greater the item difficulty parameter (b), the greater the ability level needed to reach a 50% probability of responding to an item correctly. In other words, the greater the value of b, the more difficult the item is. In IRT, the test taker's ability (θ) lies between –4 and +4 according to origin of the normal curve. This statement is an assumption underlying b values. The value of b theoretically lies between $-\infty$ and $+\infty$. According to Hambleton and Swaminathan (Retnawati, 2014, p. 17), an item is said to be good in terms of difficulty level if value of b of this item ranges between –2 and +2. When the value of b of an item is close to –2, the difficulty level of that item is very low, whereas when the value of b of an item is close to +2, the difficulty level of that item is very high.

A number of previous studies have sought to uncover the characteristics of a test and test items with a focus on the Indonesian language subject. Among these studies is one conducted by Suryani (2017). Her study focused on investigating the characteristics of Indonesian language test items under the CTT approach used in the end-of-semester examinations for tenth graders of a senior high school in a certain regency based on three aspects of analysis, namely difficulty level, differentiability, and distractor effectiveness. In addition, Anggraini and Suyata (2014) conducted a study that in addition to revealing the quality of the Indonesian language test used in national-standardized school examinations (*Ujian Akhir Sekolah Berstandar Nasional*) at the elementary school level, also aimed to reveal the characteristics of the test items based on CTT which included difficulty, differentiation, distractor effectiveness, and reliability estimation and based on modern test theory with several focuses, one of which was the level of item difficulty. Although some studies have attempted to investigate the characteristics of tests and their items with a focus on the Indonesian language subject, studies with a focus on this subject have received less attention than other subjects such as mathematics or sciences. Given the relevant existing studies, the current study aims to reveal the characteristics of the Indonesian language test items used in the USBN at the senior high school level based on the CTT and 1-PL IRT model. Although we have identified a number of studies that have a similar focus to our study in terms of subject matter, namely Indonesian language, we still believe that our study can still provide a contribution in providing additional insight or understanding in the effort to improve the quality of Indonesian language testing and learning through the context of this study which is different from previous studies.

METHOD

This was a descriptive study that focused on item analysis to investigate the characteristics of the Indonesian language test items used in the USBN in the academic year of 2018/2019. In order to reveal the characteristics of these items, data from the responses of twelfth grade students at a public senior high school in the Special Region of Yogyakarta, Indonesia, to the test were analyzed using the CTT approach and modern test theory. The test was administered to students through two test packages, namely package A and package B, where each student only responded to one of the two packages. A total of 218 students took the test, with details of 109 students responding to package A and 109 other students responding to package B. Because the Indonesian language test that is the focus of this study is part of a standardized examination, as we mentioned earlier, everything related to the development, administration, analysis of results, and evaluation of

the results of these tests has been regulated by the National Education Standards Board (2018) to ensure the standardization of these tests and examinations. This Indonesian language test consisted of 40 five-option multiple-choice items and 5 essay items; where students were given 120 minutes to respond to the items. However, in our study, we only focused on the characteristics of multiple-choice test items; or in other words, we only paid attention to the responses that students gave to the multiple-choice items contained in the test. It has also been regulated that the items contained in the test must be developed in accordance with the blueprint provided by the Ministry of Education and Culture of the Republic of Indonesia. From the blueprint, there were 20 to 25 percent of the items that served as anchor items that were ready to be used and the rest, namely 75 to 80 percent of the items needed to be developed by the education unit or school through the subject teacher forum in accordance with the existing blueprint (National Education Standards Board, 2018). The test blueprint provided information about the content scope of the test and the cognitive level being tested. The blueprint for the Indonesian language test in addition to providing these two types of information also provided information about the types of texts used as stimulus or discourse on a test item. In the test blueprint, it has been mentioned that there were three cognitive levels that were focused on in the Indonesian language test, they are knowledge and understanding, application, and reasoning. The test covered six topics, namely reading non-literary texts, reading literary texts, writing non-literary texts, writing literary texts, characteristics and structure of texts, and linguistic features of texts.

We analyzed the data of students' responses on the Indonesian language test, particularly on the part of multiple-choice items – we had permission from the school principal and the Indonesian language subject teacher to use the data – by first coding them into binary form, 1 or 0, where 1 indicates a correct response for an item and 0 for an incorrect response. We then used this binary student response data in item analysis under CTT and modern test theory approach. Item analysis was devoted to revealing item characteristics in terms of difficulty and discriminating power. Under the CTT approach, the difficulty level of a test item contained in a test package was determined based on the proportion correct (p) which expresses the proportion of the number of students who answered correctly or obtained a score of 1 on a test item to the number of students who took the test package. Based on the meaning of p, it is clear that p ranges from 0 to 1. We then used the p value of an item to categorize the difficulty level of the item into easy, moderate, or difficult. Following the guidelines commonly used in previous studies in categorizing the difficulty level of an item based on the p value (Bichi & Embong, 2018; Maharani & Putro, 2020; Rafi et al., 2023; Suryani, 2017), we declared that an item was easy when the p value of the item was more than 0.7, difficult when the p value of the item was less than 0.3, and else the item was said to be moderate in terms of difficulty level. Furthermore, the discriminating power of a test item was determined based on the point-biserial correlation coefficient (r_{pbis}) – a discrimination index representing the correlation between scores that students obtain on an item and total scores that students obtain on the part of multiple-choice items of the test, where an item with r_{pbis} more than 0 is considered good at discriminating students based on their abilities, while when it is less than 0, it means that the item is not good at discriminating student abilities. We analyzed the item characteristics under the CTT approach using AnBuso (Muhson, 2017).

In addition to revealing the characteristics of multiple-choice items on the Indonesian language test based on CTT, we also used modern test theory to reveal the characteristics of these test items. In using this modern test theory, we needed to determine which model, whether one-parameter, two-parameter, three-parameter, or four-parameter logistic models we should use that fits our data. Based on the item-fit test through chi-square test, the one-parameter logistic (1-PL) model is the most suitable model for estimating the characteristics of the Indonesian language test items. In addition, we chose the logistic parameter model by considering our sample size, the number of respondents from each test package, and the length of the test. Considering that the test length of each test package we analyzed was 40 items and the respondents in each test package was 109, it has been suggested that it was better for us to use the one-parameter logistic (1-PL)

model to reveal the characteristics of the test items in each test package (Şahin & Anıl, 2017). Since the model we used was 1-PL, the only characteristic we revealed was the difficulty of the test item (i.e., item difficulty parameter, b). We used the estimated difficulty parameter of each item as the basis for categorizing the difficulty of the item into three categories: easy when $b < -2.00$, moderate when $-2.00 \leq b \leq 2.00$, and difficult when $b > 2.00$ (Hambleton et al., 1991). We analyzed the characteristics of test items under the 1-PL model using Quest (Adams & Khoo, 1996). Last but not least, in addition to revealing the characteristics of the test items and categorizing them based on the CTT approach and the 1-PL model, we also qualitatively analyzed the test items with high difficulty, or difficult items, based on both approaches to describe the possible reasons why these items were classified as difficult. Qualitative analysis of such difficult items is expected to provide teachers with insights or understanding in identifying students' difficulties in learning or solving problems related to the topics contained in the difficult items. Ultimately, this will provide an opportunity for teachers to improve the Indonesian language learning they facilitate in the future for their students.

FINDINGS AND DISCUSSION

In this section, we provide and describe the results we have obtained in our study, namely the characteristics of multiple-choice items in the Indonesian language test used in the USBN at the senior high school level. We first present the results of the analysis in our study based on the CTT which includes the characteristics of test items based on difficulty and discriminating power. The results of the analysis based on modern test theory or item response theory (IRT) according to the 1-PL model are then presented followed by a discussion on test items that fall into the difficult category based on CTT and the 1-PL IRT model. Discussion of difficult test items based on CTT and IRT can provide opportunities for teachers to improve the quality of learning.

Characteristics of Test Items according to Classical Test Theory (CTT)

We first analyzed student response data on the two Indonesian language test packages, package A and package B, used in the USBN based on the CTT approach. Through this approach, we revealed the difficulty level and discriminating power of the items contained in each test package based on the proportion correct (p) and the point-biserial correlation coefficient (r_{pbis}), respectively. Table 1 presents the results of our analysis on the difficulty level of the multiple-choice items in the two test packages, which we categorized into easy, moderate, and difficult. Based on Table 1, package A consists of mostly easy items. This is demonstrated by 25 (62.5%) items that fall into the easy category, 11 (27.5%) items that fall into the moderate category, and 4 (10%) items that fall into the difficult category. However, when we considered that the average difficulty of its items is about 0.691, test package A then has a moderate difficulty level. Not much different from package A, although package B generally consisted of easy items as shown by 22 (55%) items in the easy category, 14 (35%) items in the moderate category, and 4 (10%) items in the difficult category, the package B is moderate in terms of difficulty because the average difficulty level of its items is about 0.69. The two test packages are thus likely to be of similar difficulty.

Table 1. Characteristics of Test Items in Package A and Package B by Difficulty Based on CTT

Test Package	Easy item ($p > 0.7$)	Moderate item ($0.3 \leq p \leq 0.7$)	Difficult item ($p < 0.3$)
Package A	1, 2, 3, 4, 5, 7, 9, 11, 12, 14, 15, 16, 17, 18, 19, 21, 22, 23, 25, 26, 30, 34, 37, 39, and 40	8, 10, 13, 20, 24, 27, 28, 29, 31, 32, and 36	6, 33, 35, and 38
Package B	1, 3, 4, 5, 6, 7, 9, 14, 15, 16, 17, 18, 20, 21, 23, 25, 26, 30, 34, 37, 39, and 40	8, 10, 11, 12, 13,19, 22, 24, 27, 28, 29, 31, 32, and 36	2, 33, 35, and 38

Table 2. Characteristics of Test Items in Package A and Package B by Discriminating Power

Test Package	Item with Good Discriminating Power	Item with Not Good Discriminating Power
Package A	1, 3, 4, 5, 8, 10, 11, 13, 14, 15, 16, 17, 18, 19, 21, 22, 23, 24, 26, 27, 28, 29, 32, 33, 34, 36, 37, 39, and 40	2, 6, 7, 9, 12, 20, 25, 30, 31, 35, and 38
Package B	2, 5, 6,7 , 9, 10, 11, 14, 15, 16, 17, 20, 24, 26, 27, 29, 30, 31, 34, 35, and 36	1, 3, 4, 8, 12, 13, 18, 19, 21, 22, 23, 25, 28, 32, 33, 37, 38, 39, and 40

Table 2 presents the analysis results of our study that focused on revealing the differentiated power of the test items contained in package A and package B. We used the point-biserial correlation coefficient (r_{pbis}) as the basis for determining whether the test items had good or not good discriminating power. Whenever an item has a negative r_{pbis}, it is considered the item with a bad discriminating power, while when the r_{pbis} of an item is positive, it is considered the item with a good discriminating power. Based on this, the majority of the items in package A have good discriminating power, with details that there are 29 (72.5%) items that have good discriminating power, while the remaining 11 (27.5%) items have not good discriminating power (see Table 2). With reference to the results presented in Table 2, in contrast to package A where the difference between items with good and not good discriminating power was very contrasting in terms of number, in package B, although the number of items with good discriminating power (i.e., 21, 52.5% items) was greater than the number of items with not good discriminating power (i.e., 19, 47.5% items), the difference between the two was very small, differing by only two items. In terms of the characteristics of the test items based on the discriminating power, it can thus be said that package A is better than package B. The large number of test items with bad discriminating power contained in the package B indicates that many students who have high test performance as indicated by their scores on the multiple-choice item section of the test tend to answer incorrectly on items with negative r_{pbis}. Such test items should be majorly revised with respect to the distractors on the item or even discarded (Costello et al., 2018; Osterlind, 1998; Rafi et al., 2023).

Characteristics of Test Items according to Item Response Theory (IRT)

In the previous section we have reported the results of our analysis of student responses to two packages of Indonesian language test in terms of difficulty and discriminating power under the CTT approach. In this section, we report the results of the analysis that we conducted to reveal the characteristics of the test items based on the IRT approach. Through the IRT approach, the characteristics of the multiple-choice items contained in the two test packages were investigated using a logistic parameter model. The decision of which logistic parameter model should be used was based on the number of items that fit the model based on the chi-square test, the length of the test, and the number of respondents in each test package. From these three considerations, the 1-PL IRT model was the most suitable to use. Through this model, the characteristics of test items are represented by their difficulty level.

Table 3. Characteristics of Test Items in Package A and Package B Based on 1-PL IRT

Test Package	Easy ($b < 2.00$)	Moderate ($-2.00 \leq b \leq 2.00$)	Difficult ($b > 2.00$)
Package A	18, 21, and 23	1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19, 20, 22, 24, 25, 26, 27, 28, 29, 30, 31, 32, 34, 36, 37, 39, and 40	6, 33, 35, and 38
Package B	18, 20, 21, and 23	1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19, 22, 24, 25, 26, 27, 28, 29, 30, 31, 32, 34, 36, 37, 39, and 40	2, 33, 35, and 38

Table 3 reports the items that fall into the easy, moderate, and difficult in terms of their difficulty level as indicated by the magnitude of the parameter b of each test item. It has been identified that package A contains more items with difficulty levels in the moderate category than items that have difficulty levels in the easy or difficult category. Based on the 1-PL IRT approach,

of the 40 multiple-choice items contained in package A, three (7.5%) items were easy items, 33 (82.5%) items were easy items, and four (10%) items were difficult items. Not much different from package A, package B is also composed of items that mostly have a moderate level of difficulty (i.e., 32, 80% items) with the number of items in the easy category equal to the number of items in the difficult category (i.e., 4, 10% items each). Based on these results, it was identified that package A and package B tended to have a similar composition of item difficulty levels. In addition, the indication of packages A and B that tend to have difficulty at the moderate level as demonstrated by the majority of the items that composed each test package are in the moderate category is supported by the average of the estimated value of *b* parameter of the test items in each packages. Our analysis revealed that the average difficulty levels of the items in packages A and B were 0.00025 and 0.0995, respectively, meaning that both packages have difficulty level in the moderate category. This is consistent with the results of our investigation into the item characteristics of the tests in both packages under the CTT approach.

Difficult Test Items Based on CTT and 1-PL IRT Model

The following are test items that are considered difficult according to classical test theory (CTT) and 1-PL IRT model. The results showed that the topics is about observational report text, job application letters, and written text. This topic is included in the difficult category in both packages, both package A and package B. This is in accordance with the observations of Mugianto et al. (2017), who said that in the teaching and learning process, especially in the basic competencies (*Kompetensi Dasar*, KD) related to producing text, there are still many students who experience difficulties. They have difficulty in producing text or writing according to the text thinking structure and language rules. The results of this study can be evaluated in making items of Indonesian language and that will be used for the future test. The following are the results of the analysis of the items that are considered difficult.

Siswa diberi kutipan teks karya tulis. *Pertanyaan:* Hubungan kalimat dalam kutipan di atas menunjukkan hubungan.... A. Syarat B. Perbandingan C. Pertentangan D. Sebab akibat E. Penambahan	Translation: Students are given an excerpt of a written work. The relationship of the sentences in the excerpt above shows.... A. Conditional conjunction B. Comparative conjunction C. Adversative conjunction D. Causal conjunction E. Additive conjunction

Figure 1. First Item that is Considered as Difficult Item Based on CTT and IRT

Based on Figure 1, students are asked to show the type of conjunction contained in the quoted text of the paper. This topic is taught in senior high school, for example in learning complex sentences. In learning complex sentences, students learn types of conjunction in complex sentences consisting of coordinating conjunctions, subordinating conjunctions, comparative conjunctions, causal conjunctions, and others. Based on the results of the study, there were 26(23.85%) students who answered this item correctly.

Based on Figure 2, students are asked to identify the use of language in the two quoted texts. Students are expected to communicate effectively through coherent texts and well-ordered sentences, including spelling and punctuation of words, sentences, and more comprehensive text (Suherli et al., 2017). Language rules for standard languages in recount text and news text have been learned by students since grade VIII. The standard language is also learned in discussion and exposition texts in grade X.

Peserta didik diberi dua kutipan teks.
Pertanyaan:
Perbedaan penggunaan bahasa pada kedua kutipan tersebut adalah....

	Teks 1	Teks 2
A.	bahasa baku	ejaan baku
B.	ejaan tidak baku	ejaan baku
C.	ejaan baku	ejaan tidak baku
D.	kata tidak baku	kata-kata baku
E.	struktur kalimat tidak baku	struktur kalimat baku

Translation:
Learners are given two text excerpts.
Question:
The difference in language use in the two excerpts is....

	Text 1	Text 2
A.	standard language	standard spelling
B.	non-standard spelling	standard spelling
C.	standardized spelling	nonstandard spelling
D.	non-standard words	standardized words
E.	non-standard sentence structure	standardized sentence structure

Figure 2. Second Item that is Considered as Difficult Item Based on CTT and IRT

The quoted text in Figure 2 is the exposition text, an argumentative text along with facts. The text aims to explain and provide information to the reader. The linguistic rules in the exposition text are standard language, standardized spelling, standardized words, and standard sentence structure. The standard language is often used to bridge the understanding of many circles. Standard languages are better understood by the public (Kosasih, 2017). Based on the study, only 23 out of 109 students (21.2%) answered the item correctly. That low number can happen because students do not master the language rules comprehensively. Therefore, in addition to considering items that must be continuously evaluated, improving the ability of students to master the language rules is also important, for example, through learning with different process approach, explicitly focused on writing skills. Students often practice creating examples of exposition texts to implement the knowledge gained.

Item in the Figure 3 discusses the job application letter topic as the exposition text. National-standardized school examinations of academic year of 2018/2019 are structured based on the 2013 Curriculum learning, which is genre-based pedagogy and content language integrated learning (CLIL). Several types of text studied in the 2013 Curriculum are report, negotiation, explanatory, exposition, descriptive, procedure, and narration (Suryaman et al., 2018). Job application letters use formal language because they are addressed to official institutions. This topic should have been taught in class XII of Indonesian senior high school, as follows basic competencies (KD) of 3.1, 3.2, 4.1, and 4.2 on the 2013 Curriculum (Suryaman et al., 2018). Based on test item in the Figure 3, students are asked to fill in the contents of the job application letter based on the given text. According to the result, no students answered correctly. Students are distracted by the options.

Based the item presented in Figure 4, students are asked to complete the observation report text structure. The results demonstrated that only 30 of 109 (27.5%) students answered the item correctly. The characteristics of the answers are implied in the text. Students' ability to read the observation report text becomes a problem, especially in depicting the structure of the observation report text. Besides, students are distracted by other alternative options.

Peserta didik diberi kutipan surat lamaran pekerjaan.
Pertanyaan :
Penulisan identitas yang tepat untuk melengkapi kalimat tersebut adalah....

A. Nama : Siska Wulandari
tempat, tanggal lahir : Jakarta, 5 Mei 1990
pendidikan : S1 Manajemen Informatika
: Perempuan
jenis kelamin : Jalan Kusumanegara 32, Yogyakarta
alamat

B. Nama : Siska Wulandari
Tempat, tanggal lahir : Jakarta, 5 Mei 1990
Pendidikan : S1 Manajemen Informatika
Alamat : Jalan Kusumanegara 32Yogyakarta

C. Nama : Siska Wulandari
tempat/tanggal lahir : Jakarta, 5 Mei 1990
pendidikan : S1 Managemen Informatika
jenis kelamin : perempuan
alamat : Jalan Kusumanegara 32, Yogyakarta

D. Nama : Siska Wulandari
tempat, tanggal lahir : Jakarta, 5 Mei 1990
pendidikan : S1 Managemen Informatika
jenis kelamin : perempuan
alamat : Jalan Kusumanegara 32, Yogyakarta

E. Nama : Siska Wulandari
Tempat, tanggal lahir : Jakarta, 5 Mei 1990
Pendidikan : S1 Manajemen Informatika
Jenis Kelamin : perempuan
Alamat : Jalan Kusumanegara 32, Yogyakarta

Translation:
Learners are given an excerpt of a job application letter.
Question:
The correct identity to complete the sentence is....

A. Name : Siska Wulandari
place, date of birth : Jakarta, May 5, 1990
education : S1 Management Informatics
gender : female
address : Kusumanegara Street 32, Yogyakarta

B. Name : Siska Wulandari
Place, date of birth : Jakarta, May 5, 1990
Education : S1 Management Informatics
Gender : Female
Address : Kusumanegara Street 32, Yogyakarta

C. Name : Siska Wulandari
place/date of birth : Jakarta, May 5, 1990
education : S1 Management Informatics
gender : female
address : Kusumanegara Street 32, Yogyakarta

D. Name : Siska Wulandari
place, date of birth : Jakarta, May 5, 1990
education : S1 Management Informatics
gender : female
address : Kusumanegara Street 32, Yogyakarta

E. Name : Siska Wulandari
Place, date of birth : Jakarta, May 5, 1990
Education : S1 Management Informatics
Gender : female
Address : Kusumanegara Street 32, Yogyakarta

Figure 3. Third Item that is Considered as Difficult Item Based on CTT and IRT

The result of those four difficult items can be a sight to educators about the students' experience solving difficult items. Teachers are expected to incorporate the findings of students' difficulties in solving difficult items into selecting effective and efficient learning models to improve teachers' pedagogical competence. Teachers can adjust the topics to be assessed, such as observation reports, job application letters, papers, and other types of text, so that the topic can be appropriately understood by students. Students should be motivated to learn and practice related to the subject matter to be tested.

Peserta didik diberi teks laporan hasil observasi.
Pertanyaan:
Struktur teks Laporan Hasil Observasi di atas tidak lengkap. Kalimat yang menunjukkan bagian yang tidak ada dalam teks tersebut beserta alasannya adalah....

A. Pernyataan umum, karena tidak ada penjelasan umum tentang objek.
B. Klasifikasi, karena tidak ada penjelasan tentang garis besar objek.
C. Deskripsi bagian, karena tidak ada penjelasan detail tentang objek.
D. Deskripsi manfaat, karena tidak ada penjelasan tentang manfaat objek.
E. Penegasan ulang, karena tidak ada kesimpulan dari pendapat penulis.

Translation:
Learners are given the text of an observation report.
Question:
The structure of the Observation Report text above is incomplete. The sentence that shows the missing part of the text and the suitable reason is....

A. General statement, because there is no general description of the object.
B. Classification, because there is no explanation about the outline of the object.
C. Part description, as there is no detailed explanation of the object.
D. Description of benefits, because there is no explanation of the benefits of the object.
E. Reaffirmation, because there is no conclusion of the author's opinion.

Figure 4. Fourth Item that is Considered as Difficult Item Based on CTT and IRT

The future studies are expected to explore the thing that has not be studied yet in our study, that is equating analysis of the two packages. In addition, it is also necessary to conduct a more detailed investigation, especially on the reliability estimation of the two packages. Future studies can also focus more on identifying the differential item functioning (DIF). This results of our study, therefore, are expected to contribute to providing the understanding of the characteristics of multiple-choice items contained in packages of Indonesian language test used in the national-standardized school examinations through providing opportunity to learn from what can be improved in the future assessments or examinations as well as providing opportunity for teacher to improve their Indonesian language learning practices.

CONCLUSION

This study seeks to contribute to uncovering the characteristics of multiple-choice items used in Indonesian language test in a national-standardized school examinations (*Ujian Sekolah Berstandar Nasional*, USBN). The results of revealing the characteristics of these test items are expected to be used as a basis in the development of Indonesian language test through reflection on what needs to be considered and improved in achieving a good measurement instrument and in improving the quality of Indonesian language learning practices. By using two approaches, namely the classical test theory (CTT) and item response theory (IRT) based on the one-parameter logistic model, our study demonstrates that the two packages of Indonesian language test used in the USBN tend to have difficulty levels in the moderate category. Under the CTT approach, although the majority of items contained in each package are in the easy category, when based on the average item difficulty, the two test packages analyzed in this study have a difficulty level in the moderate category. Under the IRT approach, the majority of items in both test packages are in the moderate category, and this is consistent when examined on the basis of the average estimated value of the parameter *b* of each item. Furthermore, although it has been revealed that package A and package B dominantly consist of items with good discriminating power, unfortunately there are also a high number of items with bad discriminating power. This finding brings consequences to the need for more attention to quality assurance of test items based on the power of these items in distinguishing test takers from their performance on the test. This attention can be given by improving the quality of the distractors on multiple-choice items. Lastly, the disclosure of difficult items based on the CTT and IRT approaches through this study can once again be used by teachers to improve the quality of Indonesian language learning practices in the future.

REFERENCES

Adams, R. J., & Khoo, S. T. (1996). *ACER Quest: The interactive test analysis system* (Version 2.1.) [Computer software]. Australian Council for Educational Research. https://research.acer.edu.au/measurement/3/

Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory*. Waveland Press.

Anggraini, D., & Suyata, P. (2014). Karakteristik soal UASBN mata pelajaran bahasa Indonesia di Daerah Istimewa Yogyakarta pada tahun pelajaran 2008/2009 [Characteristics of UASBN test on Indonesian language subjects in the Special Region of Yogyakarta in the academic year of 2008/2009]. *Jurnal Prima Edukasia, 2*(1), 57–65. https://doi.org/10.21831/jpe.v2i1.2644

Arruarte, J., Larrañaga, M., Arruarte, A., & Elorriaga, J. A. (2021). Measuring the quality of test-based exercises based on the performance of students. *International Journal of Artificial Intelligence in Education*, *31*, 585–602. https://doi.org/10.1007/s40593-020-00208-0

Bichi, A. A., & Embong, R. (2018). Evaluating the quality of Islamic civilization and Asian civilizations examination questions. *Asian People Journal*, *1*(1), 93–109.

Costello, E., Holland, J., & Kirwan, C. (2018). The future of online testing and assessment: Question quality in MOOCs. *International Journal of Educational Technology in Higher Education*, *15*(1), 1–14. https://doi.org/10.1186/s41239-018-0124-z

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement*. Prentice-Hall.

Evers, A., Muñiz, J., Hagemeister, C., Høstmælingen, A., Lindley, P., Sjöberg, A., & Bartram, D. (2013). Assessing the quality of tests: Revision of the EFPA review model. *Psicothema*, *25*(3), 283–291. https://doi.org/10.7334/psicothema2013.97

Gronlund, N. E., Miller, M. D., & Linn, R. L. (2009). *Measurement and assessment in teaching* (10th ed.). Pearson Education.

Hambleton, R. K., Swaminathan, H., & Rogers, D. J. (1991). *Fundamentals of item response theory*. Sage.

Huber, S. G., & Skedsmo, G. (2017). Standardization and assessment practices. *Educational Assessment, Evaluation and Accountability*, *29*, 1–3. https://doi.org/10.1007/s11092-017-9257-1

Kosasih, E. (2017). *Buku siswa bahasa Indonesia untuk kelas SMP/MTs kelas VIII [Indonesian language student book for junior high school/MTs class VIII]*. Kementerian Pendidikan dan Kebudayaan.

Maharani, A. V., & Putro, N. H. P. S. (2020). Item analysis of English final semester test. *Indonesian Journal of EFL and Linguistics, 5*(2), 491–504. https://doi.org/10.21462/ijefl.v5i2.302

Mardapi, D. (1999). *Estimasi kesalahan pengukuran dalam bidang pendidikan dan implikasinya pada ujian nasional [Estimation of measurement error in education and implications for national examinations]*. Universitas Negeri Yogyakarta.

Moses, T. (2017). A review of developments and applications in item analysis. In R. E. Bennett & M. von Davier (Eds.), *Advancing human assessment* (pp. 19–46). Springer. https://doi.org/10.1007/978-3-319-58689-2_2

Mugianto, M., Ridhani, A., & Arifin, S. (2017). Pengembangan perencanan pembelajaran menulis teks laporan hasil observasi model pembelajaran berbasis proyek siswa kelas X SMA [Development of lesson plan for writing observation report text using project-based learning model for grade X high school students]. *Ilmu Budaya: Jurnal Bahasa, Sastra, Seni dan Budaya*, *1*(4), 353–366. http://doi.org/10.30872/jbssb.v1i4.769

Muhson, A. (2017). *AnBuso* (Version 8.0) [Computer software].

National Education Standards Board. (2018). *Prosedur operasional standar penyelenggaraan ujian sekolah berstandar nasional [Standard operating procedure of the administration of the national-standardized school examination]*. https://bsnpindonesia.org/2018/12/bsnp-tetapkan-pos-usbn-dan-un-2019/

Nurgiyantoro, B. (2016). *Penilaian pembelajaran bahasa berbasis kompetensi [Competency-based assessment on language learning]*. BPFE Yogyakarta.

Osterlind, S. J. (1998). *Constructing test items: Multiple-choice, constructed-response, performance, and other formats* (2nd ed.). Kluwer Academic Publishers.

Pardede, T., Santoso, A., Retnawati, H., Rafi, I., Apino, E., & Rosyada, M. N. (2023). Gaining a deeper understanding of the meaning of the carelessness parameter in the 4PL IRT model and strategies for estimating it. *REID (Research and Evaluation in Education)*, *9*(1), 86–117. http://doi.org/10.21831/reid.v9i1.63230

Rafi, I., Retnawati, H., Apino, E., Hadiana, D., Lydiati, I., & Rosyada, M. N. (2023). What might be frequently overlooked is actually still beneficial: Learning from post national-standardized school examination. *Pedagogical Research*, *8*(1), em0145. https://doi.org/10.29333/pr/12657

Retnawati, H. (2014). *Teori respons butir dan penerapannya [Item response theory and its applications]*. Nuha Medika.

Retnawati, H., Kartowagiran, B., Arlinwibowo, J., & Sulistyaningsih, E. (2017). Why are the mathematics national examination items difficult and what is teachers' strategy to overcome it? *International Journal of Instruction, 10*(3), 257–276. https://doi.org/10.12973/iji.2017.10317a

Retnawati, H., Kartowagiran, B., Hadi, S., & Hidayati, K. (2011). Identifikasi kesulitan peserta didik dalam belajar matematika dan sains di sekolah dasar [Identifying learners' difficulties in learning math and science in primary schools]. *Jurnal Kependidikan, 41*(2), 162–174. https://doi.org/10.21831/jk.v41i2.1930

Şahin, A., & Anıl, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory & Practice*, *17*(1), 321–335. http://doi.org/10.12738/estp.2017.1.0270

Suherli, S., Suryaman, M., Septiaji, A., & Istiqomah, I. (2017). *Buku guru bahasa Indonesia untuk kelas SMA/MA/SMK/MAK kelas XI [Indonesian language book for teachers of SMA/MA/SMK/MAK class XI]*. Kementerian Pendidikan dan Kebudayaan.

Suryaman, M., Suherli, S., & Istiqomah, I. (2018). *Bahasa Indonesia untuk SMA/MA/SMK/MAK kelas XII [Indonesian for SMA/MA/SMK/MAK grade XII]*. Kementerian Pendidikan dan Kebudayaan.

Suryani, Y. E. (2017). Pemetaan kualitas empirik soal ujian akhir semester pada mata pelajaran bahasa indonesia sma di kabupaten Klaten [Empirical quality mapping of end-of-semester exam questions on Indonesian language subjects in senior high schools in Klaten district]. *Jurnal Penelitian dan Evaluasi Pendidikan, 21*(2), 142–152. https://doi.org/10.21831/pep.v21i2.10725