Pedagogical Research
2023, 8(1), em0145
e-ISSN: 2468-4929
https://www.pedagogicalresearch.com

MODESTUM

Research Article

OPEN ACCESS

What might be frequently overlooked is actually still beneficial: Learning from post national-standardized school examination

Ibnu Rafi [1*], Heri Retnawati [1], Ezi Apino [2], Deni Hadiana [3], Ida Lydiati [4], Munaya Nikma Rosyada [1]

[1] Department of Mathematics Education, Faculty of Mathematics and Natural Sciences, Yogyakarta State University, Special Region of Yogyakarta, INDONESIA
[2] Doctoral Program of Educational Research and Evaluation, Yogyakarta State University, Special Region of Yogyakarta, INDONESIA
[3] Research Center for Education, National Research and Innovation Agency (BRIN), Special Capital Region of Jakarta, INDONESIA
[4] SMA N 7 Yogyakarta, Special Region of Yogyakarta, INDONESIA
***Corresponding Author:** ibnurafi789@gmail.com

Citation: 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

ARTICLE INFO

Received: 25 Jun. 2022

Accepted: 17 Nov. 2022

ABSTRACT

This study describes the characteristics of the test and its items used in the national-standardized school examination by applying classical test theory and focusing on the item difficulty, item discrimination, test reliability, and distractor analysis. We analyzed response data of 191 12th graders from one of public senior high schools in Yogyakarta City, Indonesia, to the examination on the elective mathematics subject. The results showed that both multiple-choice and essay items contained in the test were at a moderate level of difficulty. The lowest item difficulty index went to the multiple-choice item where students failed in interpreting straight and dashed lines and went to the essay item that required complex intra-mathematical connections. In addition, we only found one item which was poor in distinguishing student competence. Furthermore, students' test scores on multiple-choice and essay items were reliable. Although most multiple-choice items had at least two functioning distractors, it was still found two items whose all distractors were not functioning. In this article, we provide some suggestions concerning improvement towards mathematics learning practices.

Keywords: Indonesia, elective mathematics subject, national-standardized school examination, test item analysis, test development and evaluation

INTRODUCTION

The education implementation through learning activities aims to facilitate students in developing their abilities in cognitive, psychomotor, and affective aspects. As we live in a society, the existence of education is to ensure the preservation of beliefs, values, and knowledge contained in that society and promote the curiosity, creativity, abilities of its members to become a better society (Bass, 1997). Additionally, Bhardwaj (2016) argued that education is important for people because it facilitates them to become better people with a good manner and character and be more confident in attempting and struggling to encounter difficulties to achieve their goals for a bright future. In Indonesia's case, education that is actualized in a learning process is held for facilitating students in developing their spirituality, potential, intelligence, personality, and skills (President of the Republic of Indonesia, 2003).

The aims of education, as mentioned earlier, are still general. They would be specified based on the subject matter at formal education, i.e., school. For instance, mathematics learning at school is conducted to promote students' mathematical proficiency that focuses not only on cognitive and skill or behavior aspects but also on affective aspects, including belief, attitude, value, and feeling. Mathematical proficiency comprises five strands: conceptual understanding, procedural fluency, strategic competence, adaptive reasoning, and productive disposition (National Research Council, 2001). Conceptual understanding deals with proficiency in understanding mathematical ideas or concepts and connecting one mathematical concept and another. Conceptual understanding can be strengthened through developing procedural fluency. Procedural fluency refers to proficiency in determining procedures and carrying out the procedures to solve mathematics problems effectively and efficiently. Procedural fluency is closely related to strategic competence, which refers to proficiency in solving mathematics problems by flexibly determining and formulating problem-solving approaches appropriate to the problem's context. In problem-solving activity, besides, we need the first three proficiencies, we also need adaptive reasoning. Adaptive reasoning deals with proficiency in justifying whether the contexts of problems, mathematical concepts, a connection of concepts, procedures, and strategies that are used make sense. While the first four proficiencies focus on cognitive and behavioral aspects, the last is a productive disposition, focusing on affective aspects. This last proficiency represents a tendency to see and appreciate the usefulness of

mathematics and learning mathematics and belief about self in learning mathematics. The mathematical proficiency is then integrated and adjusted with mathematics contents expected to learn based on education and grade level.

Conducting an educational assessment is one way to know whether the objectives have been achieved. Educational assessment is an essential process for obtaining information about students (Nitko & Brookhart, 2011; Reynolds et al., 2010) and identifying educational success (Retnawati et al., 2017), and at the same time, it guarantees the quality of the education (Simms & George, 2014). Information about students obtained from the assessment can be used to make decisions regarding student achievement in mastering specific competences and efforts to improve the quality of learning by identifying possible learning difficulties (Nitko & Brookhart, 2011) and misconceptions that students experience or errors that students do (Rafi & Retnawati, 2018). Assessment to obtain information related to student characteristics can be done through several techniques, one of which is a test. A test can be defined as an instrument in which standardized or non-standardized and systematic procedures are employed to collect and assign scores or qualitative labels to student performance (Allen & Yen, 1979; Nitko & Brookhart, 2011; Reynolds et al., 2010). A test, especially for a large-scale test, is said to be standardized when it is developed, "administered, scored, and interpreted in a standard manner" (Reynolds et al., 2010, p. 7) to guarantee that every student who takes the test receive the same directions, testing conditions, and test specifications (i.e., content, format, test length, scoring, or psychometric characteristics of items and test) such that an optimal accuracy, interpretation, and comparability of test scores obtained by students would be reached (American Educational Research Association et al., 2014).

A standardized test has a specific intended use and consequences for the students who take the test and for schools or institutions involved in the test. The test's intended use affects the test specifications, one of which is the format of the test questions or items. The format of test items can be divided into two types, namely selected-response items (or fixed-response items) and constructed-response items (or free-response items). The distinction between the two types of item format lies in students' freedom as test-takers in responding to the items and the type of students' competences to be assessed. In a test that employs selected-response items, students are required to select one or more responses from several provided options that they consider a solution. In contrast to selected-response items, constructed-response items require students to respond to the item in their own words based on their knowledge, understanding, and reasoning. Accordingly, it can be argued that constructed-response items tend to be appropriate for measuring students' higher-order reasoning skills, while selected-response items for measuring students' lower-order thinking skills (Katz et al., 2000). However, Katz et al. (2000) also asserted that students might use a similar problem-solving approach and reasoning procedure when solving selected-response items and solving constructed-response items.

A standardized test brings consequences for students and schools. The consequences can be direct and crucial (so-called high-stakes) or indirect and minor (so-called low-stakes). The consequences for students or schools when there is a failure in the high stakes standardized test can vary, ranging from not being able to graduate from a certain level of education, not being able to enroll in state universities, replacing educators, to reorganizing curriculum and schools (Nitko & Brookhart, 2011). It is different from the consequences of the low stakes standardized test, where students who show poor performance on the test may not receive any consequences other than obtaining a poor score. Based on the consequences of high-stakes and low-stakes standardized tests, it makes sense when Marsh et al. (2005) argued that student performance in a low-stakes standardized test is unlikely to be influenced by study habits, effort, and persistence unless the student has very high motivation to succeed on the test.

The Indonesian government held the national-standardized school examination in the academic year of 2018/2019. This examination was considered a low-stakes test because test score was not the only determinant of student graduation from a certain level of education. Education units or schools conducted this examination to determine student achievement or competence characteristics by referring to the blueprint set by the National Education Standards Board (National Education Standards Board, 2018). The blueprint comprises rules about cognitive level (i.e., knowledge and understanding, application, and reasoning) and scope of the material to be examined that teacher used through subject teacher forum as their guide for developing test including constructing test items.

An instrument in the form of a test and its items in the national-standardized school examination should have a good quality, especially in terms of their psychometric properties (e.g., validity and reliability). It was essential to guarantee that the instrument yields test scores with a minimal error of measurement and provide information about student achievement or performance accurately so that the examination result could help teachers in decision-making. Cohen and Swerdlik (2018) asserted that a test could be said to be good when the test is clear and has a minimum level of difficulty in administering, scoring, and interpreting the results; provide benefits to students as test-takers, teachers, and the society; measure precisely what it is supposed to measure; and reliable. The last two criteria are also considered for judging whether a test item is good besides the criterion that the test item should be able to discriminate between high and low achievement students on the test (Cohen & Swerdlik, 2018; Retnawati, 2016).

The quality of a test could not be separated from the quality of each test item that builds it up (Urbina, 2014). In other words, to obtain a good test quality, teachers in developing the test should ensure that each test item also has good quality. This can be obtained through item analysis. Item analysis, a part of the test development process, refers to procedures or techniques used to assess the characteristics of a test and its items theoretically (or qualitatively) based on reviewers' judgments or empirically (or quantitatively) based on students' responses to the test and evaluate their characteristics for improvement (Cohen & Swerdlik, 2018; Crocker & Algina, 2008; Retnawati, 2016; Urbina, 2014). The item analysis results from test tryout help teachers decide which items they need to include, revise, or exclude. Furthermore, Nitko and Brookhart (2011) suggest that item analysis is helpful for teachers to provide proper feedback on student performance on tests and do self-reflection on difficulties that students might

experience. Thus, item analysis is an essential process in developing tests and valuable to improve the quality of learning so that teachers should carry out such a process.

Although item analysis is vital for teachers when developing tests and trying to improve the quality of learning by evaluating test results, previous studies (e.g., Maharani & Putro, 2020; Muna et al., 2019; Retnawati et al., 2019) have revealed that teachers overlook item analysis for several reasons. A study that focuses on evaluating the implementation of the final examination system conducted by Retnawati et al. (2019) showed that the teacher had validated the test items used in the national-standardized school examination. Teachers carried out the validation through subject-teacher forums at the school level or regency/city level. However, although the National Education Standards Board (2018) has mandated schools to assign subject teachers to try out the instruments used on the examination, the test tryout was not carried out. Furthermore, the study of Retnawati et al. (2019) revealed that not all teachers confirmed the quality of the test items that students had done on the national-standardized school examination. This result indicates that using student responses to the national-standardized school examination by teachers has not been optimal.

On the other hand, it is an opportunity for teachers to take advantage of the student response data to consider improvements to learning methods to improve student learning outcomes through item analysis. Analysis of student responses to the examination provides feedback to teachers to identify the strengths and weaknesses of the learning methods they have applied so far (Quaigrain & Arhin, 2017; Talebi et al., 2013) and improve characteristic quality of items that they might use in the future examinations (Quaigrain & Arhin, 2017). Therefore, based on the willingness to obtain things that can be learned from the national-standardized school examination instrument, the present study seeks to describe characteristics of the instrument that reflect the quality or characteristics of the test and its items on the examination by performing post-examination item analysis. Although several studies focusing on item analysis of examination instruments have been carried out (e.g., Argianti & Retnawati, 2020; Sampouw & Retnawati, 2020), our study was still relevant to be conducted considering that the national-standardized school examination instrument could be regarded as a teacher-developed test. Consequently, each test instrument in a particular subject used by schools in a regency or city would provide opportunities for learning that are different from the instruments used by schools in other regencies or cities.

THEORETICAL REVIEW

Item Analysis for Test Development and Post-Examination Evaluation

Test development is typically carried out through stages, from test conceptualization, test construction, test tryout, and item analysis to test revision (Cohen & Swerdlik, 2018; Crocker & Algina, 2008; Retnawati, 2016; Urbina, 2014). Item analysis is an essential step in developing a test (Rodriguez, 2005). It plays a role in assisting teachers in revising test items that do not meet good characteristics. The revision can be in the form of adding, excluding, or modifying certain items as needed. Urbina (2014) divides item analysis into two types: qualitative item analysis and quantitative item analysis. Qualitative item analysis, usually carried out by experts in the subject area, focuses on assuring that the test item is following the objectives to be achieved by using these items, does not contain any potential bias, free from grammatical and technical errors (Crocker & Algina, 2008; Urbina, 2014), and is representative to measure the important competences that students have learned (Ebel & Frisbie, 1991). Meanwhile, quantitative item analysis, which is carried out through statistical procedures using student response data on the test tryout or the real test, emphasizes assessing the test item's level of difficulty and discriminating power (Urbina, 2014).

Although item analysis is a procedure that is part of the test development stage for examination purposes, this procedure can also be carried out after the examination. This kind of item analysis is known as post-examination analysis (Tavakol & Dennick, 2011, 2012, 2016). Post-examination analysis can be conducted after the test on the examination was administered to students as test-takers, and students' works on the test were scored. Furthermore, this analysis can be considered to improve the examination cycle (i.e., a recursive process that includes test development, administration, and evaluation) (Tavakol & Dennick, 2016) primarily aimed at improving the quality and reliability of the test and its items (Tavakol & Dennick, 2011) if they are considered for future use. In addition, Tavakol and Dennick (2011, 2016) list other functions of conducting post-examination analysis: adjusting student scores on defective items, developing item banks, and guaranteeing evidence of the validity and reliability of test scores. The benefits of conducting a post-examination analysis are in line with the assertion of Ebel and Frisbie (1991) that evaluating the characteristics of the test and its items after the examination is administered and scored provides an opportunity for teachers to learn to develop their professional competences (e.g., test item-writing skills) and optimize student learning.

Conducting item analysis to develop tests and evaluate the examinations benefits teachers and students. However, item analysis, especially quantitative item analysis, is impossible when a test tryout is also not conducted; and in practice, this may happen for several reasons. Tavakol and Dennick (2011) explained that, on the one hand, the test tryout should indeed be carried out for item analysis purposes to obtain good quality test items for examination. Still, on the other hand, the tryout test could result in leakage of examination items, so it was deemed unnecessary. As an alternative, they argue that the test items used in the examination should at least have been improved based on the suggestions for improvement that the experts gave. Thus, when quantitative item analysis is not conducted in the test development process, teachers can still benefit from the procedure through post-examination item analysis.

Many previous studies have used the classical test theory (CTT) (e.g., Argianti & Retnawati, 2020; Awopeju & Afolabi, 2016; Maharani & Putro, 2020; Muna et al., 2019; Quaigrain & Arhin, 2017; Retnawati et al., 2017; Sampouw & Retnawati, 2020) or item response theory (IRT) (e.g., Awopeju & Afolabi, 2016; Gleason, 2008) approach in performing item analysis. CTT is based on a linear relationship between an observable variable, namely the observable test score (X), and the sum of two unobservable (or latent)

variables, namely the true score (*T*) and error score (*E*). Although CTT has several shortcomings, such as the dependence of item statistics and test characteristics on students taking the test and the dependence of the characteristics of these students on the test used (Hambleton & Jones, 1993; Magno, 2009; Zanon et al., 2016), many studies still used the CTT approach for several considerations. These considerations lead to the advantages of CTT, such as CTT does not require a large sample size for item parameter estimation, and it is easy for test data to meet the linear model and assumptions of the CTT (Hambleton & Jones, 1993). Post-examination analysis in CTT focuses on item statistics such as item difficulty, item discrimination, and item distractors and test statistics such as test-score mean, standard deviation, and test score reliability. This is in line with those mentioned by Impara and Plake (Magno, 2009) that the evaluation of a test with CTT approach usually includes an analysis of the total test score, item difficulty, item-total correlation which represents item discrimination, item distractors, and test score reliability.

Item Difficulty

When a test item is scored dichotomously, for example, on multiple-choice items where a score of one is assigned for each student who answers correctly and a score of zero for those who answer incorrectly on that item, the mean score of all students on that item is the same as the proportion of students who answered the item correctly. The proportion, denoted by p_i where *i* represents the *i*-th item on the test, reflects item difficulty (or item facility) index whose value ranges from 0.00 (for a perfectly difficult item) to 1.00 (for a perfectly easy item). The concept of item difficulty as the average score of students on multiple-choice items also applies to essay items. Based on this concept, essay item difficulty is indicated by the students' average score relative to the maximum or minimum possible item score.

The closer the average score is to the maximum possible item score, the easier the item is. Essay item difficulty is reformulated as the proportion of the difference of average score for the item and minimum possible item score to the difference of maximum and minimum possible item score to equalize the item difficulty scale on multiple-choice items (Nitko & Brookhart, 2011).

There are several views regarding the interpretation of the *p*-value or item difficulty index. Retnawati et al. (2017) consider items with a *p*-value less than 0.375 as difficult items for students. Quaigrain and Arhin (2017) stated that items with a *p*-value less than 0.2 were too difficult and more than 0.9 were too easy. Quaigrain and Arhin (2017) and Reynolds et al. (2010) suggest that the best *p*-value is within the range of 0.4 and 0.6, with a mean of 0.5, because the discrimination index will be maximum in that range. Allen and Yen (1979) mentioned slightly different things, where they state that a *p*-value in the range between 0.3 and 0.7, with a mean of 0.5, is considered the optimal level of item difficulty because it will be able to provide maximum information regarding differences (ability) among students.

The optimal mean *p*-value is not always 0.5. It may also depend on a varying number of choices. Lord (Reynolds et al., 2010) suggests that the optimal mean *p*-value for five-option multiple-choice items is 0.69, while for constructed-response (e.g., essay) items is 0.5. Apart from the different interpretations of the *p*-value and the optimal mean *p*-value, it needs to be understood that the *p*-value essentially reflects a behavioral measure (Quaigrain & Arhin, 2017). In other words, the difficulty index of the test item does not necessarily represent the intrinsic characteristics (Quaigrain & Arhin, 2017) or content (Ebel & Frisbie, 1991) of the item, but it also relatively represents the ability of the group of students to respond to the item (Ebel & Frisbie, 1991; Haladyna & Rodriguez, 2013; Rodriguez, 2005) which makes it hard to be estimated accurately (Haladyna & Rodriguez, 2013).

Item Discrimination

Item discrimination or item discriminating power refers to the extent to which the test item's capacity to distinguish between students who have good competence (high-scoring students) and students who have less competence (low-scoring students) on the test. Such test item's capacity is represented by a value that ranges from -1.00 to 1.00, which is referred to as a *d*-value or item discrimination index. The common method used to determine a *d*-value of a particular multiple-choice item (dichotomous item) is by calculating the difference between the proportion of students in the high-scoring group who answered the item correctly and the proportion of students in the low-scoring group who answered the item correctly. Kelley (Crocker & Algina, 2008; Reynolds et al., 2010) and Rodriguez (2005) suggested that each high- and low-scoring group consisted of 27% of the total number of students. When it comes to an essay item (polytomous item), the item discrimination index can be expressed as the difference between the high-scoring and low-scoring group averages (Nitko & Brookhart, 2011). When the *d*-value of an item is negative, the item fails to discriminate between students who have good competence and those who have less competence. In other words, students in the high-scoring group who are supposed to answer the item correctly answered the item incorrectly and vice versa for students in the low-scoring group. Such item needs to be revised or discarded.

Some reports of multiple-choice item analysis performed by a computer program suggest that item discrimination index is typically represented by a point-biserial correlation coefficient (denoted by r_{pb}) (Ebel & Frisbie, 1991; Reynolds et al., 2010) that is mathematically equivalent to the Pearson correlation coefficient (*r*) (Allen & Yen, 1979; DiBattista & Kurzawa, 2011) for essay items. These two correlation coefficients range from -1.00 to 1.00, it is the same with the *d*-value, and can be interpreted in the same manner as the *d*-value, where a non-positive value indicates the item is unable to discriminate competent students from those who are less competent on the test. Even though the coefficient is positive, if the value is very small, the corresponding test item is still considered insufficient in distinguishing students in terms of their competence on the test. There are various opinions about how much the value of the correlation coefficient of an item so that the item can be said to have good discriminating power. Some suggest a minimum value of 0.2 (Chiavaroli & Familari, 2011; DiBattista & Kurzawa, 2011; Ding & Beichner, 2009), and some others suggest a minimum value of 0.25 (Reynolds et al., 1994) or even 0.3 (Abdel-Hameed et al., 2005; Barker & Ansorge, 2007; Osadebe, 2015; Retnawati, 2016).

Test Score Reliability

The term reliability is frequently associated with test characteristics although basically, that term reflects the degree to which students' test scores are consistent across time, content, or scorers. Therefore, as Bardhoshi and Erford (2017) asserted, it is more suitable to use the term reliability to refer to test score reliability (i.e., "test scores are reliable") rather than test reliability (i.e., "the test is reliable"). Based on the CTT, the concept of reliability is associated with the proportion between true and observed score (or test score) variances in a population of students (Bardhoshi & Erford, 2017; Cohen & Swerdlik, 2018; Matheson, 2019; Miller, 1995; Price, 2017; Tavakol & Dennick, 2012) which ranges from zero to one. The observed score variance is derived from two variances, namely the true score variance which is assumed to be systematic, and the error variance which is assumed to be random. Taking this concept into account, we can immediately notice that a reliability coefficient (or reliability index) is none other than a percentage of true score variance (Bardhoshi & Erford, 2017; Cohen & Swerdlik, 2018) (e.g., 0.7 means 70% true score variance and 30% error variance). Because what is obtained from a test is the observed score, while the true score can never be directly and with certainty determined, the reliability coefficient then can only be estimated through certain approaches depending on reliability attributes. Reliability attributes are related to how a test is administered to students or the number of test scorers.

Under certain conditions, an assessment can only be completed by students through a single administration or a single form of a test. Accordingly, the reliability of test scores can be estimated through measures of internal (or inter-item) consistency, which reflect error variance due to content sampling or item heterogeneity (Reynolds et al., 2010) and the relationship of items of a test (Bardhoshi & Erford, 2017; Reynolds et al., 2010). Coefficient alpha (α; well-known as Cronbach's alpha), which can be interpreted as "the mean of all possible split-half coefficients" (Cronbach, 1951, p. 331), is considered as one of the measures of internal consistency that is widely used because, as argued by Price (2017), it is effective to estimate score reliability of a test that comprised of dichotomously or polytomously scored items. Reliability coefficients, including coefficient alpha, have a direct relationship with the standard error of measurement (SEM), in which by knowing SEM, students' true scores can be predicted based on the confidence interval.

There are various views regarding the size of reliability coefficient so that students' test scores can be accepted or said to be reliable even though these views are mainly based on the common basis, namely the purpose of using the test or how important the decision to be made based on the test results is. It is required a high-reliability coefficient for a standardized test or a test that is used to make notable decisions, i.e., at least 0.85 (Abdel-Hameed et al., 2005; Ebel & Frisbie, 1991) or 0.90 (Bardhoshi & Erford, 2017). As for a test that of little consequence as the results of that test are not used to derive crucial decisions about students or a teacher-made assessment or classroom test, the lower reliability coefficient ranging from 0.5 to 0.7 is sufficient (Reynolds et al., 2010; Rudner & Schafer, 2002; Wells & Wollack, 2003). Even though reliability estimates of 0.8 or greater is more desirable in many testing conditions (Reynolds et al., 2010) and judged to be very good (Ursachi et al., 2015), a reliability estimate of 0.7 or greater is still believed to be reliable (Borozová & Rydval, 2014; Ding & Beichner, 2009; Tavakol & Dennick, 2012).

Distractor Analysis

One component of a multiple-choice item is response options, which consist of a keyed option and several incorrect options (or distractors). Distractors should be plausible to students who have less competence, experience misconceptions or misunderstandings, or make errors so that their frequencies of being selected by those students are more significant than the keyed option. In other words, distractors of a multiple-choice item are expected to be more chosen by students in the low-scoring group than those in the high-scoring group who are believed to have mastered the content or concepts used to answer the item correctly. An analysis of distractors has been recognized to be an essential part of item analysis (Hingorjo & Jaleel, 2012; Testa et al., 2018; Urbina, 2014) as it allows teachers to discard poorly functioning distractors and identify some possible misconceptions and misunderstandings that students have and errors that students do when they solve problems. Distractor analysis is also a means for teachers to detect the possibility of a mis-keyed item indicated by most high-scoring students choosing a particular distractor compared to the keyed option on that item (Nitko & Brookhart, 2011).

In general, there are three views regarding statistical methods used by researchers in their studies to distinguish between functioning and non-functioning distractors on a multiple-choice item of a test. The first view is only focused on the frequency of students choosing each distractor, in which a distractor is said to be functional when at least 5% of the total number of students taking the test choose that distractor (e.g., Argianti & Retnawati, 2020; Haladyna & Downing, 1993; Hingorjo & Jaleel, 2012; Sampouw & Retnawati, 2020). The second view is emphasized on the frequency of students choosing each distractor and the index or coefficient of discrimination. It is said that a distractor works adequately when not less than 5% of all students choose it and it has a negative discrimination index or coefficient (e.g., DiBattista & Kurzawa, 2011; Haladyna & Downing, 1988; Quaigrain & Arhin, 2017; Testa et al., 2018). Others (e.g., Chiavaroli & Familari, 2011; Maharani & Putro, 2020) suggested that a distractor can be considered to perform effectively when its index or coefficient of discrimination is much less than the keyed option. However, it is preferable if the value is negative. The last view is emphasized on the use of the choice mean of a distractor, which is the average test score of students choosing the distractor, to identify a non-functioning distractor (Haladyna & Downing, 1993; Haladyna & Rodriguez, 2013), where the higher choice mean of a distractor than the choice mean of keyed option, the worse the distractor (Gierl et al., 2017; Haladyna & Rodriguez, 2013).

Purpose of the Present Study

As we have mentioned earlier, the main purpose of the present study was to describe the characteristics of elective mathematics test and its items, where this test was part of national-standardized school examination. The characteristics of the test and its items were investigated based on item difficulty level and factors that possibly make certain items to have certain

difficulty level, the ability of multiple-choice items to distinguish students based on their test performance, test score reliability, and the potential of distractors to provide information of misconceptions or misunderstandings that students might have and errors that students might do in solving the test.

METHOD

This descriptive study focusing on quantitative item-analysis used response data of 191 twelfth graders from a public senior high school in Yogyakarta City, Indonesia, on the national-standardized school examination in the elective or non-compulsory mathematics subject. Elective mathematics was only for students in the mathematics and science program. The examination employed a test consisting of 30 multiple-choice (selected-response) items with five choices (i.e., A, B, C, D, and E: one keyed option and four distractors) and five essay (constructed-response) items. Students were given 120 minutes to work on the test. The Ministry of Education and Culture had provided about seven to eight items (anchor items) along with the rules for numbering the test items (National Education Standards Board, 2018). Teachers developed the remaining items through subject teacher forums under the coordination of the Provincial Education Department.

The test measured students' cognitive abilities at the level of knowledge and understanding, application, and reasoning by covering content that students have learned from tenth to twelfth grade including algebra, calculus, geometry and trigonometry, and statistics. Students were awarded a score of one for each multiple-choice item answered correctly and awarded a score of zero when incorrectly answering or not responding to that item. While in the type of essay item, students were awarded scores ranging from zero (as the minimum possible score) to eight (as the maximum possible score). Thus, on the test items in the form of essays, the minimum and maximum scores that students could receive were zero and 40, respectively. Data were available in a spreadsheet containing students' responses on multiple-choice items and scores on essay items.

We analyzed the data by applying the CTT approach and employing jMetrik (Meyer, 2014) and Microsoft Excel. We checked the accuracy of the keyed option of multiple-choice items before performing item analysis and we did not find any miskeyed item. The first focus analyzed in this study was the level of difficulty of the test items determined by considering an item difficulty index. For multiple-choice items, the item difficulty index was represented by the proportion of the number of students who correctly answered an item and the number of students who took the examination. As for essay items, the item difficulty index was firstly determined based on the mean of students' score on the corresponding item and then it was converted into an adjusted item difficulty index by following the rule proposed by Meyer (2014) and Nitko and Brookhart (2011) to equalize the scale with that on multiple-choice items. Because the maximum and minimum possible scores that students could obtain on an essay item were eight and zero, respectively, the adjusted item difficulty index of an essay item refers to the division result between the mean of student' score on that item and eight.

Considering the suggestion that has been put forward by Allen and Yen (1979) in regard to the best difficulty index, we considered an item whose difficulty index was within the range between 0.3 to 0.7 as the item with a moderate level of difficulty. Accordingly, items whose difficulty index was less than 0.3 were said to be difficult, while those whose difficulty index was greater than 0.7 were said to be easy. Such determination of item difficulty level category has also been used in several previous studies (e.g., Adegoke, 2013; Bichi and Embong, 2018). In addition to providing item difficulty index and its difficulty category, three multiple-choice items and one essay item that have a low difficulty index are provided as a means to call for improvement towards mathematics learning practices.

The second focus of this study was to describe the power of multiple-choice items in discriminating competent students who are indicated by a high score they obtained from those who are less competent because of a low score that they obtained on the multiple-choice test section. Information on how well the discriminating power of items was identified based on the item discrimination index represented by the point-biserial correlation coefficient (r_{pb}). In jMetrik, r_{pb} could be obtained by selecting Pearson correlation in the item-total correlation type panel (Meyer, 2014). In the present study, a multiple-choice item was considered to be good at discriminating students in regard to their competence on the test when its item discrimination index was not less than 0.3 as suggested by some researchers (see Abdel-Hameed et al., 2005; Barker & Ansorge, 2007; Osadebe, 2015; Retnawati, 2016).

The analysis of the item discriminating power that is focused on the keyed option has a relation to distractor analysis. It was expected that item distractors have characteristics that are opposite to keyed option. In this study, we did not only consider the percentage of students who choose a distractor but also consider the discrimination index of that distractor to include a distractor into a functioning distractor category. By following a number of previous studies (e.g., DiBattista & Kurzawa, 2011; Haladyna & Downing, 1988; Quaigrain & Arhin, 2017; Testa et al., 2018), in this study, a distractor is said to be functional when it was chosen by at least 5% of students who took the test, and its discrimination index is negative. As for test score reliability, we used Cronbach's alpha for estimating reliability, where the coefficient alpha of 0.7 is a minimum standard to say that test scores were acceptable in terms of their reliability.

RESULTS

Item Difficulty

The analysis results regarding item difficulty levels based on the proportion of students answering correctly on related multiple-choice items are presented in **Table 1**. **Table 1** has demonstrated that item difficulty index within the range of 0.168 to

Table 1. Item difficulty of multiple-choice test items

Item	Item difficulty index (category)	Item	Item difficulty index (category)
1	0.937 (easy)	16	0.189 (difficult)
2	0.738 (easy)	17	0.335 (moderate)
3	0.895 (easy)	18	0.487 (moderate)
4	0.712 (easy)	19	0.796 (easy)
5	0.838 (easy)	20	0.476 (moderate)
6	0.639 (moderate)	21	0.382 (moderate)
7	0.173 (difficult)	22	0.471 (moderate)
8	0.393 (moderate)	23	0.780 (easy)
9	0.168 (difficult)	24	0.277 (difficult)
10	0.822 (easy)	25	0.539 (moderate)
11	0.869 (easy)	26	0.456 (moderate)
12	0.262 (difficult)	27	0.319 (moderate)
13	0.838 (easy)	28	0.796 (easy)
14	0.466 (moderate)	29	0.251 (difficult)
15	0.288 (difficult)	30	0.435 (moderate)

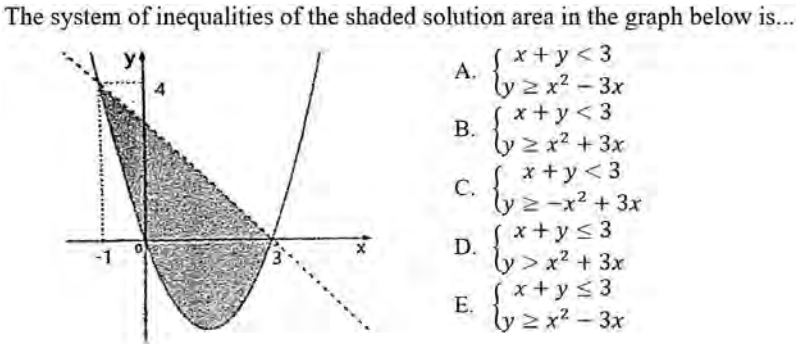


Figure 1. The difficult multiple-choice item related to the system of inequalities in two variables (linear-quadratic) (Retrieved from the main package of school examinations in the academic year 2018/2019 for the Mathematics and Natural Sciences Program in Mathematics (Elective or Non-Compulsory) subject)

0.937. Furthermore, of 30 multiple-choice items, there were seven difficult items, 12 moderate items, and 11 easy items. Of the seven difficult items, item 9 was identified as the most difficult because only 16.8% of students correctly answered this item. Furthermore, the analysis result from jMetrik showed that all items' mean level of difficulty or the proportion of students who answered correctly for all items was 0.534 (SD=0.248). Overall, the multiple-choice items contained in the test were at a moderate level of difficulty.

The first difficult item was item 7 that related to a system of inequalities in two variables (**Figure 1**) and required students to determine the system of inequalities representing the shaded area bounded by a straight line and a parabola. The first step that students should do in determining the proper system of inequalities was to determine the equations of a straight line and a parabola. This step might be challenging for students. But then, the provided options helped students in determining these equations. By taking the provided options into account, it was easy for students to conclude that the equation for the straight line is $x+y=3$. Afterwards, students just needed to decide which quadratic equation that represents the parabola; $y=x^2-3x$, $y=-x^2+3x$, or $y=x^2+3x$. In deciding this one, students needed to have a sufficient conceptual and principal understanding as well as mathematical connection skills in making the interconnection between the coefficient of x^2 and the direction in which a parabola opens and interconnection between the points passed through by the parabola and quadratic equation which represents that parabola. Students who had such understanding and skills would easily conclude that $y=-x^2+3x$ did not represent the given parabola. Then they just needed to determine the equation by examining which equation, either $y=x^2-3x$ or $y=x^2+3x$, satisfied by $x=3$ and $y=0$. The intended quadratic equation is $y=x^2-3x$. The next challenging step was to assign the proper inequality signs. The analysis result from jMetrik showed that most students (44.5%) chose option E, followed by those who chose option A (17.3%).

Next, this study revealed that the most difficult item is item 9, which is related to determining the integration by parts of an indefinite integral (**Figure 2**). Students should recognize the integration method and understand when they should use that method to solve this problem. The rule of integration by parts is $\int udv=uv-\int vdu$. However, the students frequently abandon this rule for practical reasons and prefer to use the tabular integration by parts (see Alcantara, 2015; Horowitz, 1990). Through this method, the students just need to differentiate the function $u(x)=2x+5$ twice to get zero and integrate another function $v(x)=(4x+1)^{1/2}$ twice. The integration of $f(x)=(ax+b)^r$, where r is a rational number could become a challenging task for the student, especially for students who struggle to perform fraction operations. Through tabular integration by parts, the students would obtain the integration results as follows: $\int(2x+5)(4x+1)^{1/2}dx=(2x+5)((4x+1)^{3/2}/6)-2((4x+1)^{5/2}/60)$.

The result of $\int (2x+5)\sqrt{4x+1}dx$ is…

A. $\frac{1}{5}(4x+1)^{\frac{3}{2}}(x+4)$
B. $\frac{1}{5}(4x+1)^{\frac{5}{2}}(x+4)$
C. $\frac{1}{6}(4x+1)^{\frac{3}{2}}(x+4)$
D. $\frac{1}{6}(4x+1)^{\frac{5}{2}}(x+4)$
E. $\frac{1}{6}(4x+1)(x+4)$

Figure 2. The difficult multiple-choice item related to the integration by parts (Retrieved from the main package of school examinations in the academic year 2018/2019 for the Mathematics and Natural Sciences Program in Mathematics (Elective or Non-Compulsory) subject)

The size of the pupil of an animal's eye is expressed by the formula $f(x) = (30x^{-0.2} + 40)/(3x^{-0.2} + 5)$ where f in mm dan x represents the intensity of light received by the pupil of the eye. The size of the pupil of the eye when the light received is of great intensity is … mm

A. 2
B. 4
C. 6
D. 8
E. 10

Figure 3. The difficult multiple-choice item related to the limit at infinity (Retrieved from the main package of school examinations in the academic year 2018/2019 for the Mathematics and Natural Sciences Program in Mathematics (Elective or Non-Compulsory) subject)

Table 2. Item difficulty of essay test items

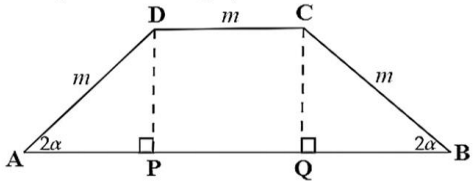
Item	Item difficulty index (*p*)	Adjusted item difficulty index (*p**)	Category of difficulty
31	5.262	0.658	Moderate
32	4.073	0.509	Moderate
33	6.408	0.801	Easy
34	4.414	0.552	Moderate
35	4.398	0.550	Moderate

Unfortunately, to arrive at the answer key, that is $(1/5)(4x+1)^{3/2}(x+4)$, the students were required to manipulate or simplify that algebraic expression by factorizing. Based on the output of analysis by using jMetrik, we uncovered the information that only 16.8% of students who chose the answer key, while nearly half of the subject of this study (49.7%) chose option C. These 49.7% of students who took the examination might have understood that one of the common factors of $(2x+5)((4x+1)^{3/2}/6)$ and $2((4x+1)^{5/2}/60)$ is $(4x+1)^{3/2}$, but they failed in doing algebraic manipulation in the expression of $((2x+5)/6)-(2(4x+1)^{1/2}/60)$.

The third difficult item, item 12, deals with the connection between mathematics and another discipline according to the context used (**Figure 3**). In this item, students were asked to determine the size of pupils of an animal, represented by $f(x)$, where x expresses the light intensity received by pupils when the light intensity is enormous. The mathematics concept contained in this item is the concept of limits at infinity, i.e., the limit of $f(x)$ as x approaches infinity. Because the analysis results showed that the answers concentrated on the options E (50.8%) and D (26.2%), we can say that the students have understood the mathematics concept behind this word problem. However, because the majority of the students chose the option E, it indicated that there were still a lot of students who had an understanding that the limit of $f(x)=(a_nx^n+a_{n-1}x^{n-1}+...+a_0)/(b_mx^m+b_{m-1}x^{m-1}+...+b_0)$ as x approaches to infinity is a_n/b_m for the case of $n=m$. They were unaware and might not fully understand what happened to the limit of $f(x)$ when n is a negative number and x approaches infinity. The low number of students who could correctly answer the item 12 indicated that students lack knowledge about basic concepts of limit at infinity of quotient of two functions. Students might only notice that the value of limit of $f(x)=(a_nx^n+a_{n-1}x^{n-1}+...+a_0)/(b_mx^m+b_{m-1}x^{m-1}+...+b_0)$ as x approaches to infinity can be equal to a_n/b_m when $n=m$, ∞ or $-\infty$ when $n>m$, or 0 when $n<m$, without having sufficient understanding of the concepts behind those possibilities of that limit value.

The essay item difficulty level is demonstrated by the item difficulty index that represents the mean of students' score on that item and the adjusted item difficulty index that represents the conversion result of the item difficulty index into a continuous scale ranging from zero to one. **Table 2** shows that the item difficulty index of item 33 was 6.408, meaning that the mean of scores that students obtained on that item was 6.408. In addition, **Table 2** demonstrates that of the five essay items contained in the test, four items have a difficulty level in the moderate category. The other item had a difficulty level in the easy category when the five test

The figure below is a trapezoid ABCD with AD = CD = BC = m cm, and ∠DAB = ∠CBA = 2α (α is a cute angle).



What is the maximum area of the trapezoid (in m cm^2)? Please write down its step-by-step solution!

Figure 4. The essay item with the lowest item difficulty index (Retrieved from the main package of school examinations in the academic year 2018/2019 for the Mathematics and Natural Sciences Program in Mathematics (Elective or Non-Compulsory) subject)

Table 3. Item discrimination

Item	Discrimination coefficient	Item	Discrimination coefficient	Item	Discrimination coefficient
1	0.088	11	0.329	21	0.330
2	0.334	12	0.097	22	0.157
3	0.344	13	0.368	23	0.332
4	0.233	14	0.172	24	0.327
5	0.446	15	0.338	25	0.311
6	0.198	16	0.015	26	0.323
7	-0.001	17	0.055	27	0.206
8	0.260	18	0.252	28	0.309
9	0.217	19	0.343	29	0.062
10	0.349	20	0.437	30	0.408

items were administered to 191 students. Overall, the essay items on the test have a moderate level of difficulty (M_p=4.911, SD_p=0.946 or M_{p^*}=0.614, SD_{p^*}=0.118). Item 33 was the easiest because, on average, students could obtain 80.1% of the maximum possible score range for the essay items, i.e., eight. Item 32 was the most difficult item among the four other essay items because, on average, the students were only able to obtain 50.9% of the maximum possible score range for the essay items (**Table 2**).

The essay item which has the lowest item difficulty index and was categorized as an item with a moderate level of difficulty is that related to the connection between the concepts of trigonometry, geometry, and application of the first derivative (**Figure 4**).

In order to succeed in solving this problem, students need to comprehend the trigonometry ratio in a right triangle, the formula for the area of a trapezoid, the application of the first derivative to determine the maximum value of a function, trigonometric equations, and double-angle identity as well as know the trigonometric values of special angles. On average, students who took the test could only obtain about half of the possible maximum score that students could obtain on this item. In other words, most students could only obtain a score of four out of eight on this item. Therefore, it can be said that this problem is so complicated to be solved by students because it required students to make intra-mathematical connections.

Item Discrimination

The discrimination coefficient of the 30 multiple-choice items ranged from -0.001 to 0.446 with the mean of discrimination coefficient being 0.534 (SD=0.248) (see **Table 3**). Of the 30 multiple-choice items, item 7 was the only item with a negative discrimination coefficient, meaning that the item is not able to distinguish between students who have good competence and those who have less competence on the test. Because the minimum standard used to say that items have good discriminating power is 0.3, it means that there were only 16 (53.33%) multiple-choice items that have good power to distinguish students based on their competence on solving the problems contained in the test.

Test Score Reliability

The coefficient alphas that demonstrate test score reliability on multiple-choice and essay test items were 0.740 (95% CI [0.6833, 0.7899], SEM=2.285) and 0.703 (95% CI [0.6302, 0.7643], SEM=4.756) respectively. The coefficient alpha of 0.740 means 74% true score variance that represents consistency and 26% error variance that represents the heterogeneity of the multiple-choice items. The coefficient alpha of 0.703 can be interpreted as 70.3% true score variance that represents consistency and 29.7% error variance that represents inconsistency of essay items contained in the test. Both coefficients indicate that students' test scores were reliable or considered acceptable (≥0.70), where the coefficient alpha on multiple-choice items is slightly greater than the coefficient alpha on essay items.

Distractor Analysis

The mean number of functioning distractors per item was 2.833 (SD=1.289). Based on **Table 4**, we could say that the majority of distractors were functioning to distract students who have no sufficient competences so that they tend to choose the distractors

Table 4. Item distractor performance

Characteristics	n	%
Number of multiple-choice items	30	
Number of distractors assessed	120	
Distractor with:		
Frequency <5%	34	28.33
Discrimination coefficient ≥0	3	2.5
Frequency <5% and discrimination coefficient ≥0	2	1.67
Frequency=0%	1	0.83
Functioning distractors per test	85	70.83
Functioning distractors per item		
None	2	6.67
One	4	13.33
Two	3	10
Three	9	30
Four	12	40

rather than the keyed option. Of the 30 multiple-choice items, two items have no functioning distractors because the distractors were chosen by less than 5% of students who took the examination even though the discrimination index of all these distractors was already negative. Furthermore, the distractor analysis that has been performed shows that most multiple-choice items have three or four distractors that performed well as they should be. Moreover, 35 (29.17%) of distractors were not functioning.

DISCUSSION

The first focus of this study was to describe how difficult the test administered at the national-standardized school examination for the case of the elective mathematics subject is. This focus was not only motivated by the potential of post-examination analysis but also encouraged by the statement of Wiliam (2001) that students' scores on a test would provide more meaningful information when they are supported by information about how difficult the test is. Considering the latter, our study revealed that the test used in the examination was at a moderate level of difficulty. In the results section, we have presented three multiple-choice items that were in the difficult category and one essay item that was in the moderate category as a means to provide insight for learning into students' understanding and possible errors that students did.

From the results of the analysis on item 7 (see **Figure 1**), it can be argued that most students have had a sufficient understanding of determining the proper equations for the straight line and parabola as well as assigning the inequality sign. However, they did not fully understand or were not aware of what does the dashed line on the graph meant. Although previous studies have pointed out that students considered problems presented in the form of a graph or symbol are easier to solve rather than those presented in a common form of problem, i.e., word problem form that requires a lot of comprehensions (Arsaythamby & Julinamary, 2015), our study showed that students were still struggling in solving mathematics problem involving graphs. These results indicate a lack of student competence in representing graphs and symbols to mathematical statements. That competence of flexibly using and making multiple mathematical representations has been widely recognized for its critical role to succeed in problem-solving and the means to explain mathematical situations to others (Heinze et al., 2009).

Although item 9 (see **Figure 2**) presents a straightforward problem, the problem was challenging for students because they were required to identify which technique the best, integration by substitution or by parts, according to the integrand. When students could identify the integration by parts as the best fit technique, they need to choose which function should be $u(x)$ and which should be dv. Afterward, they needed to integrate vdu. The process to arrive at the desired solution to that problem was complicated. The complexity of the process was coupled with the need for students to perform algebraic manipulations in order to obtain algebraic forms corresponding to the available options. A number of previous studies (e.g., Borji & Font, 2019; Borji et al., 2021) have revealed that when solving integration by parts, students found it difficult to determine $u(x)$ and dv such that the determination could make it easier for them to determine $v(x)$ and $\int v\mathrm{d}u$. Other than that, Li et al. (2017) have found that techniques determination was one of the difficulties experienced by students in solving integral problems, and students were often overwhelmed with many techniques of integration. In addition, Kiat (2005) have demonstrated that insufficient content knowledge in algebra leads to technical errors that prevent students to obtain a correct answer to an integral problem. It has also been asserted by Muzangwa and Chifamba (2012) that basic algebra comprehension is important for students as a means for understanding the topics of calculus and to avoid misconceptions and make errors when learning and solving calculus including integral.

Based on the previous study performed by Sampouw and Retnawati (2020), we discovered that item 32 was an anchor item for the subject of elective or non-compulsory mathematics that has been provided by the Ministry of Education and Culture. The present study showed that item 32, which required students to make mathematical connections became the most difficult among the other four essay items. A previous study conducted by Yusron et al. (2020) has shown similar results that problems that require mathematical connections became the most difficult among items contained in tests used in national-standardized school examinations. Jailani et al. (2020) also found that most senior high school students were still struggling in making connections between mathematical representations, concepts, and procedures. In addition, Sampouw and Retnawati (2020) found two interesting results through their study. Firstly, they revealed that item 32 became an essay item with the lowest item difficulty index, which is the same as the result of our study. Secondly, they found that the adjusted item difficulty index for item 32 was 0.1,

meaning that this item was categorized as a difficult item because, on average, most students could only obtain 10% of the possible maximum score on that item due to students' low mastery of competences with respect to the application of the first derivative. On the basis of the same standard which was used to categorize the level of item difficulty, the latter study result of Sampouw and Retnawati (2020) is contrastive from what we found in our study.

The second focus of our study was to describe the extent to which multiple-choice items contained in the test could distinguish competent and incompetent students in terms of their score on the multiple-choice section of the test through their discrimination coefficient. With respect to this focus, the results of our study revealed that more than half of all multiple-choice items performed well in distinguishing students who took the test based on their score assumed to reflect their competence. The previous study focusing on investigating the characteristics of a mathematics test and its items used in the national-standardized school examination (Argianti & Retnawati, 2020) has also shown similar results in terms of the number of items that have good discriminating power. Furthermore, a study conducted by Sampouw and Retnawati (2020) which has a similar focus to our study obtained a similar result with what we have found that of 30 multiple-choice items, one item has a negative discrimination coefficient. In our study, the item with a negative discrimination coefficient was the most difficult one based on its difficulty index. Nitko and Brookhart (2011) suggested that when many students from high-scoring group incorrectly answered an item, we need to assure whether the answer key of the item is correct. As we have mentioned earlier, there is no mis-keyed item. Accordingly, as affirmed by Nitko and Brookhart (2011), this result just reflects the lack of student competences required to solve the problem.

The present study has demonstrated that students' test score on the multiple-choice and essay sections were reliable indicated by coefficient alpha which is greater than 0.7. Reliability of test scores is important not only because of its role as pre-requisite for test item validity (Reynolds et al., 2010; Wells & Wollack, 2003; Wiliam, 2001) which reflects the accuracy of justification for student competence (Wells & Wollack, 2003) but also because it can indicate random measurement error contained in the student's test score (Wells & Wollack, 2003). Such type of error that makes student's test score unpredictable can arise due to students, test-specific, or scoring specific factors (Wells & Wollack, 2003). In a study conducted by Argianti and Retnawati (2020), the students' scores obtained from a test used in national-standardized school examination was reliable, meaning that this result is consistent with our result. In another study (Sampouw & Retnawati, 2020), however, showed a contrast result, in which it was found that the scores obtained from the test used in the examination was unreliable. The comparison towards the results of other studies has indicated that the reliability of test scores on the national-standardized school examination was varied. The use of multiple-choice items in a test brings consequences on the need to pay attention to the quality of distractors. Our study disclosed that nearly a third of all distractors were not functioning as they infrequently chosen by students or have negative discrimination coefficient and the mean number of functioning distractors per item was 2.833.

Implications for Mathematics Learning Practice

This study has disclosed a number of multiple-choice and essay items contained in the test that were categorized as difficult based on difficulty index and possible reasons why those items have such characteristics. These results raise several suggestions that teachers can consider improving the quality of mathematics learning that they facilitate. Because the most difficult item on the test is likely caused by students' carelessness in interpreting the dashed line on the graph that represents a linear-quadratic system of inequalities, we suggest teachers engage students in a learning activity that enable them to use multiple mathematical representations, especially from graphical representation to mathematical statements. This suggestion could also be directed to others mathematics contents such as determining the solution of linear programming based on the shaded area and inequalities based on the use of open and closed dots on a number line. The second most difficult item was the one about the limit of the quotient of two functions at infinity believed to be caused by insufficient knowledge about the limit at infinity that the limit of $1/x$ as x approaches to infinity is zero. Students may have been trapped by the practical formula to solve routine problems about the limit of the quotient of two functions at infinity. Accordingly, in addition to providing practical formulas, teachers should encourage students to derive the practical formulas by themselves based on their understanding and reasoning of the related basic concepts.

The investigation on test item difficulty has also indicated the struggle of students in solving problems that required them to make and use mathematical connections, especially intra-mathematical connections. Because García-García and Dolores-Flores (2018) argued that there is a strong relationship between understanding and connections, and it was supported by the results of study performed by Jailani et al. (2020), we encourage teachers to focus more on promoting students' mathematical understanding. Taking the results of previous study which found that unfamiliarity towards problems that require mathematical connections leads them to have difficulties in solving the problems (Jailani et al., 2020) into account, the teachers are suggested to present mathematical and application problems, in which through such problems and the process of finding the solutions to the problems, students are facilitated to make intra-mathematical connections (García-García & Dolores-Flores, 2020).

Limitations and Future Study Direction

We believe that what we have done could still be improved through future studies by considering some limitations of our study. Firstly, we could not have a detailed test blueprint used by teachers to develop test items, so we could not investigate and provide the evidence of content validity of test items used in the examination. Accordingly, if it is possible to obtain access and permission to use a test blueprint, the further study could add the investigation on the content validity as conducted by previous studies (e.g., Argianti & Retnawati, 2020; Sampouw & Retnawati, 2020). It is hoped that the investigation would not only focus on the extent to which the validity of the test items, but it should also be focused on the reason that make an item must be revised or omitted from a test due to its validity issue. The results of such investigation could help (mathematics) teachers and related policy makers in improving the ability of teachers for constructing better test items for a classroom or larger scale assessment. Other than that, a

series of trainings for teachers to be acquainted with and practice the issues of creating tests and their items, get awareness of test properties, and evaluation of tests.

Secondly, the present study only employed a response data on the national-standardized school examination of students in one school in one city only. The future study is suggested to perform item analysis on the response data of students from at least two schools or two different regencies/cities, so it is possible to make a comparison for creating the better quality of the test, learning, and student achievement. Lastly, although it is possible to use a more robust approach such as Rasch analysis for estimating item parameter considering the adequacy of the sample size (see Chen et al., 2014), in this study, we decided to just use CTT approach. Accordingly, when the sample size is adequate, the future study is directed to employ the more robust approach including Rasch analysis or even IRT for estimating item and ability parameter based on data of students' responses on the national-standardized school examination.

CONCLUSION

This study has demonstrated the benefits of preforming an analysis on a test and its items used in the national-standardized school examination. Although this examination was considered as a summative assessment which was mainly focused to evaluate student learning or attainment of basic competences at the end of student's study period at the high school level, through post-examination analysis that we have performed, the results of analysis provide benefit for teachers to create better learning process and assessment in the future. A better learning process could be attained by taking the results regarding item difficulty index and distractor analysis into account. The difficulty category of test items used in the examination varied, especially for multiple-choice items, where it could be found three categories of difficulty, i.e., easy, moderate, and difficult. Meanwhile, for the case of essay items, we only found two categories of difficulty, i.e., easy and moderate. Performing distractor analysis in this study, besides being able to be used to evaluate the quality of multiple-choice item distractor, has helped us in identifying some possible students' misunderstandings and difficulties in solving mathematics problems. It has been argued that the ability of students in making and using mathematical representations, doing algebraic manipulation, performing integration by parts, understanding basic concepts that can be used to derive a practical formula, and creating mathematical connections needs to be more encouraged.

Improving the quality of a test planned to be administered in an examination is usually done by considering the results of the analysis of the test tryout data. Due to several reasons including concerns about leakage of test items and time barriers, the improvement towards the quality of test may not be optimal. Performing post-examination analysis is considered as an alternative way for teachers to develop a better test in the future in regard to the quality of the difficulty and discriminating power of test item, test score reliability, and distractor functioning. Our study has indicated that a number of test items administered in national-standardized school examination need to be revised or omitted because of its discrimination index. This way could increase the test score reliability besides could improve the ability of test items in distinguishing students based on their competence on the examination. Moreover, the present study still found that almost a third of all distractors were not functioning, implying that in the future test development, teachers are expected to create more plausible distractors.

Author contributions: All authors have sufficiently contributed to the study and agreed with the results and conclusions.

Funding: No funding source is reported for this study.

Ethical statement: Authors stated that the study used the documents or materials collected from a completed research project (Evaluation of system of final examination in Indonesia) in 2019 conducted by the second and fourth authors of this article and their colleagues under a collaborative research scheme between Yogyakarta State University and Research and Development Agency, Ministry of Education and Culture, Indonesia. The collaborative project includes several activities, including analyzing test characteristics in a number of subjects including Mathematics (Elective or Non-compulsory) used in national-standardized school examinations in a number of senior high schools in three provinces in Indonesia (i.e., Special Region of Yogyakarta, South Kalimantan, and East Nusa Tenggara). Authors further stated that the researchers in this collaborative project have obtained approval from the Provincial Education Department and the schools concerned to use the examination documents for use in research projects to improve the quality of learning and educational assessment in Indonesia in the future.

Declaration of interest: No conflict of interest is declared by authors.

Data sharing statement: Data supporting the findings and conclusions are available upon request from the corresponding author.

REFERENCES

Abdel-Hameed, A. A., Al-Faris, E. A., Alorainy, I. A., & Al-Rukban, M. O. (2005). The criteria and analysis of good multiple choice questions in a health professional setting. *Saudi Medical Journal, 26*(10), 1505-1510.

Adegoke, B. A. (2013). Comparison of item statistics of physics achievement test using classical test and item response theory frameworks. *Journal of Education and Practice, 4*(22), 87-96.

Alcantara, E. C. (2015). On the derivation of some reduction formula through tabular integration by parts. *Asia Pacific Journal of Multidisciplinary Research, 3*(1), 80-84.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Brooks/Cole.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Argianti, A., & Retnawati, H. (2020). Characteristics of math national-standardized school exam test items in junior high school: What must be considered? *Jurnal Penelitian dan Evaluasi Pendidikan* [*Journal of Educational Research and Evaluation*], *24*(2), 156-165. https://doi.org/10.21831/pep.v24i2.32547

Arsaythamby, V., & Julinamary, P. (2015). Students' perception on difficulties of symbols, graphs and problem solving in economic. *Procedia-Social and Behavioral Sciences, 177*(1), 240-245. https://doi.org/10.1016/j.sbspro.2015.02.401

Awopeju, O. A., & Afolabi, E. R. I. (2016). Comparative analysis of classical test theory and item response theory based item parameter estimates of senior school certificate mathematics examination. *European Scientific Journal, 12*(28), 263-284. https://doi.org/10.19044/esj.2016.v12n28p263

Bardhoshi, G., & Erford, B. T. (2017). Processes and procedures for estimating score reliability and precision. *Measurement and Evaluation in Counseling and Development, 50*(4), 256-263. https://doi.org/10.1080/07481756.2017.1388680

Barker, B. S., & Ansorge, J. (2007). Robotics as means to increase achievement scores in an informal learning environment. *Journal of Research on Technology in Education, 39*(3), 229-243. https://doi.org/10.1080/15391523.2007.10782481

Bass, R. V. (1997). The purpose of education. *Educational Forum, 61*(2), 128-132. https://doi.org/10.1080/00131729709335242

Bhardwaj, A. (2016). Importance of education in human life: A holistic approach. *International Journal of Science and Consciousness, 2*(2), 23-28.

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

Borji, V., & Font, V. (2019). Exploring students' understanding of integration by parts: A combined use of APOS and OSA. *EURASIA Journal of Mathematics, Science and Technology Education, 15*(7), 1-13. https://doi.org/10.29333/ejmste/106166

Borji, V., Radmehr, F., & Font, V. (2021). The impact of procedural and conceptual teaching on students' mathematical performance over time. *International Journal of Mathematical Education in Science and Technology, 52*(3), 404-426. https://doi.org/10.1080/0020739X.2019.1688404

Borozová, H., & Rydval, J. (2014). Analysis of exam results of the subject 'applied mathematics for IT.' *Journal on Efficiency and Responsibility in Education and Science, 7*(3-4), 59-65. https://doi.org/10.7160/eriesj.2014.070303

Chen, W. H., Lenderking, W., Jin, Y., Wyrwich, K. W., Gelhorn, H., & Revicki, D. A. (2014). Is Rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? An example using PROMIS pain behavior item bank data. *Quality of Life Research, 23*(2), 485-493. https://doi.org/10.1007/s11136-013-0487-5

Chiavaroli, N., & Familari, M. (2011). When majority doesn't rule: The use of discrimination indices to improve the quality of MCQs. *Bioscience Education, 17*(1), 1-7. https://doi.org/10.3108/beej.17.8

Cohen, R. J., & Swerdlik, M. E. (2018). *Psychological testing and assessment: An introduction to tests and measurement*. McGraw-Hill Education.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334. https://doi.org/10.1007/BF02310555

DiBattista, D., & Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests. *The Canadian Journal for the Scholarship of Teaching and Learning, 2*(2), 1-23. https://doi.org/10.5206/cjsotl-rcacea.2011.2.4

Ding, L., & Beichner, R. (2009). Approaches to data analysis of multiple-choice questions. *Physical Review Special Topics-Physics Education Research, 5*(2), 1-17. https://doi.org/10.1103/PhysRevSTPER.5.020103

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

García-García, J., & Dolores-Flores, C. (2018). Intra-mathematical connections made by high school students in performing calculus tasks. *International Journal of Mathematical Education in Science and Technology, 49*(2), 227-252. https://doi.org/10.1080/0020739X.2017.1355994

García-García, J., & Dolores-Flores, C. (2020). Exploring pre-university students' mathematical connections when solving calculus application problems. *International Journal of Mathematical Education in Science and Technology, 51*(7), 1-25. https://doi.org/10.1080/0020739X.2020.1729429

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research, 87*(6), 1082-1116. https://doi.org/10.3102/0034654317726529

Gleason, J. (2008). An evaluation of mathematics competitions using item response theory. *Notices of the AMS, 55*(1), 8-15.

Haladyna, T. M., & Downing, S. M. (1988). Functional distractors: Implications for test-item writing and test design. In *Proceedings of the Annual Meeting of the American Educational Research Association* (pp. 1-20).

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement, 53*(4), 999-1010. https://doi.org/10.1177/0013164493053004013

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge. https://doi.org/10.4324/9780203850381

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38-47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Heinze, A., Star, J. R., & Verschaffel, L. (2009). Flexible and adaptive use of strategies and representations in mathematics education. *ZDM-International Journal on Mathematics Education, 41*(5), 535-540. https://doi.org/10.1007/s11858-009-0214-4

Hingorjo, M. R., & Jaleel, F. (2012). Analysis of one-best MCQs: The difficulty index, discrimination index and distractor efficiency. *Journal of the Pakistan Medical Association, 62*(2), 142-147.

Horowitz, D. (1990). Tabular integration by parts. *The College Mathematics Journal, 21*(4), 307-311. https://doi.org/10.1080/07468342.1990.11973325

Jailani, J., Retnawati, H., Apino, E., & Santoso, A. (2020). High school students' difficulties in making mathematical connections when solving problems. *International Journal of Learning, Teaching and Educational Research, 19*(8), 255-277. https://doi.org/10.26803/ijlter.19.8.14

Katz, I. R., Bennett, R. E., & Berger, A. E. (2000). Effects of response format on difficulty of SAT-mathematics items: It's not the strategy. *Journal of Educational Measurement, 37*(1), 39-57. https://doi.org/10.1111/j.1745-3984.2000.tb01075.x

Kiat, S. E. (2005). Analysis of students' difficulties in solving integration problems. *The Mathematics Educator, 9*(1), 39-59.

Li, V. L., Julaihi, N. H., & Eng, T. H. (2017). Misconceptions and errors in learning integral calculus. *Asian Journal of University Education, 13*(1), 17-39.

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *The International Journal of Educational and Psychological Assessment, 1*(1), 1-11.

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

Marsh, H. W., Köller, O., Trautwein, U., Lüdtke, O., & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering. *Child Development, 76*(2), 397-416. https://doi.org/10.1111/j.1467-8624.2005.00853.x

Matheson, G. J. (2019). We need to talk about reliability: Making better use of test-retest studies for study design and interpretation. *PeerJ, 7*(1), 1-25. https://doi.org/10.7717/peerj.6918

Meyer, J. P. (2014). *Applied measurement with jMetrik*. Routledge. https://doi.org/10.4324/9780203115190

Miller, M. B. (1995). Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 2*(3), 255-273. https://doi.org/10.1080/10705519509540013

Muna, W., Hanafi, H., & Rahim, A. (2019). Analisis kualitas tes buatan guru mata pelajaran Bahasa Indonesia pada siswa SMP kelas IX berbasis HOTS [Analysis of the quality of tests made by Indonesian language teachers for grade IX junior high school students based on HOTS]. *Jurnal Pendidikan Bahasa* [*Journal of Language Education*]*, 8*(2), 29-40.

Muzangwa, J., & Chifamba, P. (2012). Analysis of errors and misconceptions in the learning of calculus by undergraduate students. *Acta Didactica Napocensia, 5*(2), 1-10.

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

National Research Council. (2001). *Adding it up: Helping children learn mathematics*. National Academy Press.

Nitko, A. J., & Brookhart, S. M. (2011). *Educational assessment of students*. Pearson.

Osadebe, P. U. (2015). Construction of valid and reliable test for assessment of students. *Journal of Education and Practice, 6*(1), 51-56.

President of the Republic of Indonesia. (2003). *Act of the Republic of Indonesia number 20 year 2003 on national education system*. https://www.ilo.org/dyn/natlex/docs/ELECTRONIC/84435/93875/F8347727/IDN84435.pdf

Price, L. R. (2017). *Psychometric methods: Theory into practice*. The Guilford Press.

Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education, 4*(1), 1-11. http://doi.org/10.1080/2331186X.2017.1301013

Rafi, I., & Retnawati, H. (2018). What are the common errors made by students in solving logarithm problems? *Journal of Physics: Conference Series, 1097*(1), 1-9. https://doi.org/10.1088/1742-6596/1097/1/012157

Retnawati, H. (2016). *Analisis kuantitatif instrumen penelitian* [*Quantitative analysis of research instrument*]. Parama Publishing.

Retnawati, H., Hadi, S., Munadi, S., Hadiana, D., & Muhardis, M. (2019). *Evaluasi penyelenggaraan sistem ujian akhir Indonesia* [*Evaluation of system of final examination in Indonesia*]. http://staffnew.uny.ac.id/upload/132255129/penelitian/Lap%20Akhir-Evaluasi%20Sistem%20US%20USBN%20dan%20UN%20Heri%20Retnawati%20dkk%20UNY%2010%20November%202019.pdf

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

Reynolds, C. R., Livingston, R. B., & Willson, V. (2010). *Measurement and assessment in education*. Pearson.

Reynolds, T., Perkins, K., & Brutten, S. (1994). A comparative item analysis study of a language testing instrument. *Language Testing, 11*(1), 1-13. https://doi.org/10.1177/026553229401100102

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24*(2), 3-13. https://doi.org/10.1111/j.1745-3992.2005.00006.x

Rudner, L. M., & Schafer, W. D. (Eds.). (2002). *What teachers need to know about assessment*. National Education Association of the United States.

Sampouw, F., & Retnawati, H. (2020). Characteristics of non-compulsory mathematics test items on nationally standardized school examination in Kaimana Regency, West Papua Indonesia. *Journal of Physics: Conference Series, 1581*(1), 1-8. https://doi.org/10.1088/1742-6596/1581/1/012034

Simms, M., & George, B. (2014). Approaching assessment from a learning perspective: Elevating assessment beyond technique. *Educational Assessment, Evaluation and Accountability, 26*, 95-104. https://doi.org/10.1007/s11092-013-9176-8

Talebi, G. A., Ghaffari, R., Eskandarzadeh, E., & Oskouei, A. E. (2013). Item analysis an effective tool for assessing exam quality, designing appropriate exam and determining weakness in teaching. *Research and Development in Medical Education, 2*(2), 69-72. https://doi.org/10.5681/rdme.2013.016

Tavakol, M., & Dennick, R. (2011). Post-examination analysis of objective tests. *Medical Teacher, 33*(6), 447-458. https://doi.org/10.3109/0142159X.2011.564682

Tavakol, M., & Dennick, R. (2012). Post-examination interpretation of objective test data: Monitoring and improving the quality of high-stakes examinations: AMEE Guide No. 66. *Medical Teacher, 34*(3), 161-175. https://doi.org/10.3109/0142159X.2012.651178

Tavakol, M., & Dennick, R. (2016). Postexamination analysis: A means of improving the exam cycle. *Academic Medicine, 91*(9), 1324. https://doi.org/10.1097/ACM.0000000000001220

Testa, S., Toscano, A., & Rosato, R. (2018). Distractor efficiency in an item pool for a statistics classroom exam: Assessing its relation with item cognitive level classified according to Bloom's taxonomy. *Frontiers in Psychology, 9*(1), 1-12. https://doi.org/10.3389/fpsyg.2018.01585

Urbina, S. (2014). *Essentials of psychological testing*. John Wiley & Sons.

Ursachi, G., Horodnic, I. A., & Zait, A. (2015). How reliable are measurement scales? External factors with indirect influence on reliability estimators. *Procedia-Economics and Finance, 20*(1), 679-686. https://doi.org/10.1016/s2212-5671(15)00123-9

Wells, C. S., & Wollack, J. A. (2003). *An instructor's guide to understanding test reliability*. https://testing.wisc.edu/Reliability.pdf

Wiliam, D. (2001). Reliability, validity, and all that jazz. *Education 3-13: International Journal of Primary, Elementary and Early Years Education, 29*(3), 17-21. https://doi.org/10.1080/03004270185200311

Yusron, E., Retnawati, H., & Rafi, I. (2020). Bagaimana hasil penyetaraan paket tes USBN pada mata pelajaran matematika dengan teori respons butir? [How are the results of the equating of test packages of mathematics USBN with item response theory?] *Jurnal Riset Pendidikan Matematika* [*Journal of Mathematics Education Research*], *7*(1), 1-12. https://doi.org/10.21831/jrpm.v7i1.31221

Zanon, C., Hutz, C. S., Yoo, H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexao e Critica* [*Psychology: Reflection and Criticism*], *29*(1), 1-10. https://doi.org/10.1186/s41155-016-0040-x