Jurnal Pendidikan Administrasi Perkantoran (JPAP)
Journal homepage: https://journal.unesa.ac.id/index.php/jpap

Jurnal Pendidikan Administrasi Perkantoran (JPAP)
Volume 12 Nomor 1 2024
E-ISSN: 23389621

CLASSICAL THEORY TEST: ANALYSIS OF HIGH ORDER THINKING SKILLS INSTRUMENT TEST

Agnes Dwi Anggraeni[a], Febrika Yogie Hermanto[b]

[a1] Universitas Negeri Surabaya, Indonesia
[b] Universitas Negeri Surabaya, Indonesia

ARTICLE INFO

Keywords:
Test Instrument, High Order Thinking Skills, Vocational High Schools Students, Classical Theory Test.

Article History:
Received 20 July 2024
Accepted 3 September 2024
Available online 3 October 2024

10.26740/jpap.v12n1.p194-206

ABSTRACT

This research aims to add a bank of HOTS test instruments to the elements of simple financial management for the MPLB class XI SMK major. The method used in this development is 4D, which consists of definition, design, development, and dissemination. The analysis used is Classical Test Theory (CTT), which consists of validity, reliability, discriminating power, level of difficulty, and distractor effectiveness. Test instrument using SPSS Statistic 22 and Anbuso analysis tools. As a result of the research, 30 test instruments were declared valid, and 10 test instruments were declared invalid, where the number of test instruments tested was 40 items in multiple-choice form. The 30 items of the test instrument were declared reliable with a Cornbach's Alpha value of 0.805. The discriminating power of items of test instruments shows that 22 test instruments have "good" criteria and 8 items have "fairly good" criteria. The distractor effectiveness of each item of the test instrument is declared effective with each alternative answer chosen by the student, proving that alternative answers can distract students from the correct answer. Item of test instruments that meet the criteria in CTT can be used to measure students' abilities in learning outcomes carried out at school. Moreover, test instruments with high-order thinking skills (HOTS) criteria can improve and familiarize students with higher-level thinking when solving tasks and problems while studying at school and working in the industry.

INTRODUCTION

Assessment plays a vital role in determining the quality of education and direction of learning (Pantiwati, 2015). Appropriate assessments can encourage students to improve their learning achievements (Pantiwati, 2016). The teacher's ability to carry out assessments and evaluations is needed to determine the achievement of learning objectives. Apart from that, these abilities can be used to improve the quality of the learning process carried out by teachers (Serevina et al., 2019). The learning process can be successful if the teacher can create students who can think at a high level (Desilva et al., 2020), where high-level thinking abilities are also known as High Order Thinking Skills (HOTS).

One of the problems currently faced by education providers is the low quality of teacher assessment instruments in measuring students' cognitive abilities in *High Order Thinking Skills* (HOTS) (Serevina et al., 2019). This results in students' low ability to think at a higher level, where the results of the 2015

[1] **Correspondence:**

Agnes Dwi Anggraeni, S1 Office Administration Study, Faculty of Economic and Bussiness, State University of Surabaya, Surabaya, Indonesia. Email: agnesdwi.20006@mhs.unesa.ac.id.

Program for International Student Assessment (PISA) test stated that Indonesian students were ranked 64th out of 70 countries (Permana, 2018). Furthermore, (Lestari et al., 2018) describe the PISA 2015 results scores with the title "PISA 2015 Results in Focus" in Table 1.

Table 1.
PISA SCORES 2015

	Score	Rating
Mathematics	386	63 of 72 countries
Science	403	62 of 72 countries
Read	397	64 of 72 countries

PISA is an international-level survey that aims to evaluate education worldwide by testing students' knowledge and skills, where the PISA results become an illustration of student learning outcomes throughout the world to improve the education system with better and increased teacher academic capacity and student achievement (Fenanlampir et al ., 2019).

In research conducted by (Fenanlampir et al., 2019), it is stated that the leading cause of students' failure to obtain good learning outcomes is the characteristics of the school and the learning methods or processes applied in schools, as is the case in developing countries in Southeast Asia, namely Indonesia and Colombia. Argina et al., 2017) explain that the education system in Indonesia still focuses on the formulation of science itself rather than considering the context of its application in society. For example, in the learning process, teachers are accustomed to emphasizing learning how to create data and apply formulas rather than developing critical thinking skills or training students in solving problems, where these abilities are by the characteristics of contextual assessment in HOTS developed by (Kemendikbud, 2017).

Furthermore, the observations at SMK PGRI 13 Surabaya show that the school has implemented HOTS-based test instruments in Mid-Semester and Final Semester Assessment activities. However, the number of HOTS test instruments still needs to be increased. In the MPLB Department, teachers still have difficulty creating stimuli when compiling HOTS-based test instruments, so they tend to take and modify test instruments from books, which results in students quickly guessing the answers to the test instruments given. Based on the analysis of the Odd Semester Final Assessment test instruments for the 2023/2024 Academic Year in the MPKK MPLB subjects (Office Management and Business Services Skills Concentration Subjects), it shows that of the 30 test instruments, only 3 test instruments are included in the HOTS category. In the learning process at school, students are accustomed to memorizing the concepts of material given by the teacher. They are less accustomed to developing reasoning when applying the concepts taught, so students are less able to create high-level thinking abilities. HOTS-based test instruments can encourage students to be able to interpret, analyze, manipulate, and store new information so that it can be used to solve the problems they face (Putri & Pahlevi, 2021). Therefore, the quality of HOTS-based test instruments used by Indonesian teachers must be improved.

A quality test instrument provides accurate information about students who have mastered the material and students who have not learned the material (Supranoto, 2012). To be used as a good evaluation tool, an instrument must go through the item of test instrument analysis stage to determine its suitability. Approaches that can be used include validity, reliability, discriminating power, level of difficulty, and distractor effectiveness (Muhson, 1979).

LITERATURE REVIEW AND HYPOTHESES DEVELOPMENT

Analysis of Test Instrument Items Using Classical Theory

Analyzing unlucky points using classical theory can be called classical pure score theory. Classical test theory is a fundamental theory about measuring mental abilities, which is described by the relationship between observed scores and unobserved actual scores on tests (Wang & Osterlind, 2013). (Cappelleri et al., 2014) added, classical test theory is a conventional quantitative approach to testing the reliability and validity of a scale based on its items. Classical test theory is often used to guide analyzing an instrument being developed (Sumaryanta, 2021). According to (Bichi, 2015), (Wu et al., 2016), the main parameters analyzed using classical theory are validity, level of difficulty, discriminating power, and instrument reliability. In line with this, (Magno, 2009) states that other parameters can use classical test theory, namely distractor function. (Muhson, 1979) , (Susanto et al., 2015) also argue that to determine the test quality can be measured by validity and reliability; other criteria that can be used are level of difficulty, discriminating power, and effectiveness of distractors. Validity is the ability of a test to measure what it wants to measure (Azwar, 2012). According to Mehrens & Lehmann (Azwar, 2012), reliability is the consistency between two measurement results on the same object. Discriminating power is an item of test instrument parameter used to determine whether a test instrument can differentiate between testees who have met the criteria and those who have not. The level of difficulty is one of the quality parameters of test instrument items; where if a test instrument has a level of difficulty index that is close to 0 (very difficult) or 1 (very easy), then the test instrument needs to be discarded (Azwar, 2012). The effectiveness of distractors is used to analyze test instruments in the form of multiple choices, where in the test instrument form, there are 3 to 5 alternative answers, one of which is the answer key, so the other answers must be able to distract students.

High Order Thinking Skills (HOTS)

High-level thinking is a student's thinking process at a high cognitive level, developed from various concepts and methods and by learning taxonomies (Sofyan, 2019). (Umami et al., 2021) added that high-level thinking skills are methods or techniques students use to analyze, plan, design, implement, and evaluate existing problems using their abilities. High-level thinking skills include the ability to analyze, evaluate, and create. According to Kratwohl (Purbaningrum, 2017), indicators measuring high-level thinking abilities include the following:

\

Table 2.
HOTS INDICATOR

Indicator		
Analyze	1.	Analyze incoming information and divide or structure the information into simpler parts to recognize existing patterns or relationships.
	2.	Able to recognize and differentiate the cause and effect factors of a complex scenario.
	3.	Identify/formulate questions.
Evaluating	1.	Assess solutions, ideas, and methodologies using appropriate criteria or existing standards to ensure their effectiveness or benefits.
	2.	Make hypotheses, criticize, and carry out testing.
	3.	Accept or reject a statement based on predetermined criteria
Creating	1.	Generalize an idea or way of looking at something.
	2.	Design a way to solve the problem.
	3.	Organizing elements or parts into a new structure that has never existed before.

Learning Outcomes in Simple Financial Management

The Office Management and Business Services Skills Program at Vocational High Schools equips students with the skills, knowledge, and attitudes to gain expertise in office administration management. The MPLB (Office Management and Business Services) Department is divided into phases: Phase E for class X and Phase F for classes XI and XII. The learning outcomes in element 7 of simple financial management are that students can manage petty cash, make petty cash reports, carry out simple banking transactions, and carry out cash and non-cash transactions. Thus, the test instrument indicators used to create the HOTS Test Instrument can be seen in Table 3.

Table 3.
MPLB PHASE F LEARNING ACHIEVEMENTS

No	Indicators
1	Manage petty cash
2	Make petty cash reports
3	Perform simple banking transactions
4	Carrying out cash and non-cash transactions

This research focuses on element 7 of simple financial management with learning outcomes for petty cash management and petty cash reporting because the availability of HOTS-based test instruments in this material is still limited.

METHOD

This research uses the *Research and Development* (R&D) method with Thiagarajan's 4D model, which consists of 4 stages: definition, design, development, and dissemination (Thiagarajan, 1974). However, this research is limited to analyzing test instrument items using the Classical Theory Test (CTT).). This research data is from answers from 40 HOTS test instruments with a multiple choice model tested on 56 students at SMK PGRI 13 Surabaya. Next, the data was analyzed using the Anbuso and SPSS Statistic 22 applications to reveal the analysis of the test instrument items with CTT, where the analysis started from testing 1) validity, 2) reliability, 3) discriminating power of the test instrument, 4) level of difficulty of the test instrument, and 5) distractor effectiveness (Ali, 2019). Then, validity analysis was carried out using the point biserial correlation technique with the formula (Susanto et al., 2015) :

$$r_{pbi}=\frac{M_{p-}M_t}{SD_t}\sqrt{\frac{p}{q}}$$

Information :
r_{pbi} = biserial correlation coefficient
M_p = the average score of subjects who answered correctly and whose validity is sought
M_t = average total score
SD_t = standard deviation of the total score of the proportion
p = proportion of students who answered correctly to the total number of students
q = proportion of students who answered incorrectly (q = 1-p)

The value r_{pbi} will be compared with the correlation coefficient table of the "r" *product moment value* at a significance level of 5%. If r_{pbi} the correlation coefficient result is greater (>) than the value r_{tabel}, then the results are significant, meaning the test instrument items are declared valid. The instrument is valid in the SPSS output, with a corrected item-total correlation coefficient t ≥ r_{tabel}.

The reliability analysis used in this research uses the model (Sudijono, 2013), where the reliability testing formula is as follows.

$$r_{11}=\left(\frac{n}{n-1}\right)\left(\frac{S2-\sum pq}{S2}\right)$$

Information :

r_{11} = overall test reliability
n = number of test instruments
p = proportion of students answering the test instrument correctly
q = proportion of students answering the test instrument incorrectly
$\sum pq$ = number of products of p and q
S = standard deviation

The coefficient value (r_i) will be compared with the table correlation coefficient r_{tabel}= $r_{(\alpha,n-2)}$.If r_i> r_{tabel}, then the instrument is reliable. In the SPSS output, if Cronbach's Alpha > r_{tabel}, then the instrument is reliable. A test instrument can be good if it consistently provides data from reality (Arikunto, 2013).

Discriminating power analysis was carried out using the following formula (Susanto et al., 2015) :

$$D = P_A - P_B$$

Information :

D = discriminating power index
P_A = total score in the upper group
P_B = total score in the lower group

For the record, if the number of *testees* is 100 or more, then only 27% *of the testees* in the upper group and 27% in the lower group will be used (Sudijono, 2013). Table 2 shows the interpretation of the results of discriminating power according to Saccuzzo (2009).

Table 4.
INTERPRETATION OF THE DISCRIMINATING POWER INDEX TEST ITEMS

Category	Criteria
Good	> 0.3
Pretty good	0.2 – 0.29
Not good	< 0.2

In the Anbuso application, the discriminating power analysis can be seen in the discriminating power column containing the coefficient value and the test instrument's discriminating power criteria. Discriminating power functions to improve the quality of test instrument items through empirical data and to find out how far test instrument items can measure students' ability to understand the material (Susanto et al., 2015).

Analysis of the level of difficulty of the test instrument is carried out using the following formula (Susanto et al., 2015) :

$$I = \frac{B}{J}$$

Information :

I = item difficulty index
B = proportion of students who answered correctly
J = number of students who took the test

Table 5 interprets the difficulty level of test instrument items using criteria from Saccuzzo (2009).

Table 5.
INTERPRETATION OF LEVEL OF DIFFICULTY TEST ITEMS

Intervals	Interpretation
0.00 – 0.30	Hard
0.31 – 0.70	Currently
0.71 – 1.00	Hard

A level of difficulty analysis is carried out to determine whether the test instrument is difficult, medium, or easy. A good test instrument is neither difficult nor easy (Arikunto, 2013).

(Arikunto, 2013) states that a distractor can be said to function well if it is chosen by at least 5% of students for each alternative answer. In the Anbuso application, the distractor analysis of the test instrument is displayed as a percentage of answers in the Distribution menu. The effectiveness of distractors is used as the basis for reviewing test instruments to determine whether the answers provided function as distractors (Hutabarat, 2009).

RESULTS AND DISCUSSIONS

The test instrument trial in this research was conducted by inviting students to answer the test instrument using the Quizizz application. The aim was to get responses from each respondent according to their abilities. Test instruments were developed and tested using Indonesian. Apart from that, researchers motivated respondents by rewarding respondents who answered the most correctly, starting from 1st to third most. The respondent subjects were class XI students majoring in MPLB at SMK PGRI 13 Surabaya, with 56 students consisting of 8 men and 48 women.

Instrument Validity Test

Referring to the results of the analysis of 40 test instrument items using the SPSS application, the following results were obtained:

Table 6.
VALIDITY ANALYSIS RESULTS

Test Instrument Number Code	r-count	>/<	r table	Information	Test Instrument Number Code	r-count	>/<	r table	Information
1.	0.422	>	0.2586	Valid	21.	0.225	<	0.2586	Invalid
2.	0.362	>	0.2586	Valid	22.	0.285	>	0.2586	Valid
3.	0.372	>	0.2586	Valid	23.	0.422	>	0.2586	Valid
4.	0.186	<	0.2586	Invalid	24.	0.044	<	0.2586	Invalid
5.	0.292	>	0.2586	Valid	25.	0.423	>	0.2586	Valid
6.	0,373	>	0,2586	Valid	26.	0,235	<	0,2586	Tidak Valid
7.	0,327	>	0,2586	Valid	27.	0,492	>	0,2586	Valid
8.	0,362	>	0,2586	Valid	28.	0,485	>	0,2586	Valid
9.	0,359	>	0.2586	Valid	29.	0.131	<	0.2586	Invalid
10.	0.143	<	0.2586	Invalid	30.	0.428	>	0.2586	Valid
11.	0.361	>	0.2586	Valid	31.	0.234	<	0.2586	Invalid
12.	0.433	>	0.2586	Valid	32.	0.283	>	0.2586	Valid
13.	0.443	>	0.2586	Valid	33.	0.241	<	0.2586	Invalid
14.	-0.083	<	0.2586	Invalid	34.	0.366	>	0.2586	Valid
15.	0.418	>	0.2586	Valid	35.	0,312	>	0,2586	Valid
16.	0,301	>	0,2586	Valid	36.	0,433	>	0,2586	Valid

17.	0,383	>	0,2586	Valid	37.	0,379	>	0,2586	Valid
18.	0,395	>	0,2586	Valid	38.	0,262	>	0.2586	Select
19.	0.373	>	0.2586	Select	39.	-0.190	<	0.2586	Tidak Valid
20.	0.326	>	0.2586	Select	40.	0.355	>	0.2586	Select

From the analysis data, it can be concluded that of the 40 test instruments, there were 30 test instruments (75%) which were declared valid, namely test instrument item number 1, 2, 3, 5, 6, 7, 8, 9, 11, 12, 13, 15, 16, 17, 18, 19, 20, 22, 23, 25, 27, 28, 30, 32, 34, 35, 36, 37, 38, 40. There were 10 test instruments (25%) declared invalid, namely test instrument item number 4, 10, 14, 21, 24, 26, 29, 31, 33, 39. To obtain good quality test instruments, all test instruments declared invalid will not be used in the reliability, discriminating test power, level of difficulty of the test instrument, and distractor.

Reliability Test Instruments

This research used the SPSS Statistic 22 application for reliability analysis on 30 test instruments that were declared valid. The reliability analysis found that Cornbach's Alpha was 0.805, so the instrument was declared reliable.

Discriminating power of Test Instruments

Discriminating power analysis is carried out on test instruments declared valid. Table 7 shows the results of the discriminating power analysis using the Anbuso application.

Table 7.
RESULTS OF ANALYSIS OF DISCRIMINATING POWER TEST ITEMS

Test Instrument Number Code	Coefficient	Information	Test Instrument Number Code	Coefficient	Information
1	0.388	Good	19	0.357	Good
2	0.359	Good	20	0.233	Pretty good
3	0.359	Good	22	0.244	Pretty good
5	0.253	Pretty good	23	0.392	Good
6	0.403	Good	25	0.444	Good
7	0.247	Pretty good	27	0.470	Good
8	0.344	Good	28	0.450	Good
9	0.282	Good	30	0350	Good
11	0.412	Pretty good	32	0.269	Pretty good
12	0.505	Good	34	0.393	Good
13	0.403	Good	35	0.369	Good
15	0.374	Good	36	0.471	Good
16	0.262	Pretty good	37	0.396	Good
17	0.361	Good	38	0.270	Pretty good
18	0.409	Good	40	0.355	Good

From the results of the discriminating power analysis, it can be concluded that 22 test instruments (73%) were declared "good," and 8 test instruments (27%) were declared "fairly good." Test instrument items with good discriminating power are numbered 1, 2, 3, 5, 7, 8, 10, 11, 12, 14, 15, 16, 19, 20, 21, 22, 23, 25, 26, 27, 28, 30. Meanwhile, test instrument items with good discriminating power are numbered 4, 5, 9, 13, 17, 18, 24, and 29. The decision on sound and not good discriminating power can be used because, in these two categories, items of test instruments can differentiate between students who master the material and those who do not.

Level of difficulty

Based on the results of the level of difficulty analysis of the 30 test instrument items in the Anbuso application, the following results were obtained:

Table 8.
RESULTS OF LEVEL OF DIFFICULTY ANALYSIS OF TEST ITEMS

Test Instrument Number Code	Coefficient	Information	Test Instrument Number Code	Coefficient	Information
1	0.304	Currently	19	0.125	Difficult
2	0.339	Currently	20	0.500	Currently
3	0.500	Currently	22	0.625	Currently
5	0.161	Difficult	23	0.411	Currently
6	0.125	Difficult	25	0.357	Currently
7	0.196	Difficult	27	0.464	Currently
8	0.339	Currently	28	0.143	Difficult
9	0.196	Difficult	30	0.286	Difficult
11	0.232	Difficult	32	0.286	Difficult
12	0.518	Currently	34	0.464	Currently
13	0.304	Currently	35	0.232	Difficult
15	0.536	Currently	36	0.250	Difficult
16	0.375	Currently	37	0.125	Difficult
17	0.214	Difficult	38	0.500	Currently
18	0.357	Currently	40	0.625	Currently

From the results of the level of difficulty analysis, it can be concluded that 15 test instruments (50%) were declared "difficult," and 15 test instruments (50%) were declared "medium." Test instrument items with difficulty level are found at numbers 4, 5, 6, 8, 9, 14, 16, 17, 18, 19, 25, 26, 27, 29, 30. Meanwhile, test instrument items with a level of Medium difficulty are found in numbers 1, 2, 3, 7, 10, 11, 12, 13, 15, 20, 21, 22, 23, 24, 28. That also conveys the same decision for difficult and medium test instruments as to what (the aim is to create test instrument packages with various difficulty levels).

Distractor effectiveness analysis

Based on the results of the Level of Difficulty analysis on 30 items of test instruments with the Anbuso application, the following results were obtained:

Table 9.
RESULTS OF DISTRACTOR EFFECTIVENESS ANALYSIS

Test Instrument Number Code	Spread	Functioning (Yes/No)	Test Instrument Number Code	Spread	Functioning (Yes/No)
1	Answer A: 32.1% Answer B: 7.1% Answer C: 30.4% Answer D: 10.7% Answer E: 19.6%	Yes, because all distractors work well	19	Answer A: 12.5% Answer B: 28.6% Answer C: 16.1% Answer D: 35.7% Answer E: 7.1%	Yes, because all distractors work well

2	Answer A: 10.7% Answer B: 39.9% Answer C: 12.5% Answer D: 16.1% Answer E: 26.8%	Yes, because all distractors work well	20	Answer A: 28.6% Answer B: 50% Answer C: 16.1% Answer D: 1.8% Answer E: 3.6%	Yes, because some distractors work well
3	Answer A: 21.4% Answer B: 8.9% Answer C: 14.3% Answer D: 50% Answer E: 5.4%	Yes, because all distractors work well	22	Answer A: 12.5% Answer B: 12.5% Answer C: 62.5% Answer D: 5.4% Answer E: 7.1%	Yes, because all distractors work well
5	Answer A: 19.6% Answer B: 23.2% Answer C: 23.2% Answer D: 17.9% Answer E: 16.1%	Yes, because all distractors work well	23	Answer A: 3.6% Answer B: 23.2% Answer C: 23.2% Answer D: 8.9% Answer E: 41.1%	Yes, because some distractors work well
6	Answer A: 10.7% Answer B: 17.9% Answer C: 12.5% Answer D: 41.1% Answer E: 17.9%	Yes, because all distractors work well	25	Answer A: 35.7% Answer B: 12.5% Answer C: 17.9% Answer D: 21.4% Answer E: 12.5%	Yes, because all distractors work well
7	Answer A: 39.3% Answer B: 19.6% Answer C: 8.9% Answer D: 17.9% Answer E: 14.3%	Yes, because all distractors work well	27	Answer A: 30.4% Answer B: 8.9% Answer C: 46.4% Answer D: 7.1% Answer E: 7.1%	Yes, because all distractors work well
8	Answer A: 5.4% Answer B: 14.3% Answer C: 42.9% Answer D: 3.6% Answer E: 33.9%	Yes, because some distractors work well	28	Answer A: 19.6% Answer B: 37.5% Answer C: 5.4% Answer D: 14.3% Answer E: 23.2%	Yes, because all distractors work well
9	Answer A: 7.1% Answer B: 19.6% Answer C: 7.1% Answer D: 25% Answer E: 41.1%	Yes, because all distractors work well	30	Answer A: 28.6% Answer B: 30.4% Answer C: 8.9% Answer D: 7.1% Answer E: 25%	Yes, because all distractors work well
11	Answer A: 16.1% Answer B: 23.2% Answer C: 35.7% Answer D: 12.5% Answer E: 12.5%	Yes, because all distractors work well	32	Answer A: 23.2% Answer B: 12.5% Answer C: 28.6% Answer D: 28.6% Answer E: 7.1%	Yes, because all distractors work well
12	Answer A: 17.9% Answer B: 51.8% Answer C: 7.1% Answer D: 16.1% Answer E: 7.1%	Yes, because all distractors work well	34	Answer A: 32.1% Answer B: 3.6% Answer C: 46.6% Answer D: 8.9% Answer E: 8.9%	Yes, because all distractors work well
13	Answer A: 25% Answer B: 17.9% Answer C: 12.5% Answer D: 14.3% Answer E: 30.4%	Yes, because all distractors work well	35	Answer A: 16.1% Answer B: 26.8% Answer C: 30.4% Answer D: 3.6% Answer E: 23.2%	Yes, because some distractors work well
15	Answer A: 21.4% Answer B: 1.8% Answer C: 5.4% Answer D: 17.9% Answer E: 53.6%	Yes, because some distractors work well	36	Answer A: 5.4% Answer B: 16.1% Answer C: 25% Answer D: 51.8% Answer E: 1.8%	Yes, because some distractors work well

16	Answer A: 19.6% Answer B: 17.9% Answer C: 37.5% Answer D: 10.7% Answer E: 14.3%	Yes, because all distractors work well	37	Answer A: 12.5% Answer B: 28.6% Answer C: 16.1% Answer D: 35.7% Answer E: 7.1%	Yes, because all distractors work well
17	Answer A: 12.5% Answer B: 10.7% Answer C: 33.9% Answer D: 21.4% Answer E: 21.4%	Yes, because all distractors work well	38	Answer A: 28.6% Answer B: 50% Answer C: 16.1% Answer D: 1.8% Answer E: 3.6%	Yes, because some distractors work well
18	Answer A: 10.7% Answer B: 23.2% Answer C: 35.7% Answer D: 16.1% Answer E: 14.3%	Yes, because all distractors work well	40	Answer A: 12.5% Answer B: 12.5% Answer C: 62.5% Answer D: 5.4% Answer E: 7.1%	Yes, because all distractors work well

The results of the level of difficulty analysis show that 23 test instruments (76%) have answer keys in the effective category, and 7 test instruments (24%) have answer keys in the ineffective category. Test instrument items that have effective distractor effectiveness are test instruments number 1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 13, 14, 15, 16, 18, 19, 21, 23, 24, 25, 26, 27, 28. Meanwhile, the test instrument items with distractor effectiveness could be more effective, namely test instrument numbers 7, 12, 17, 20, 22, 29, and 30.

Implications for Practice

Question validity provides an empirical picture of the test's quality (Dachliyani, 2019). Test instruments that have been declared valid will be used to analyze reliability, discriminating power, level of difficulty, and distractor effectiveness, while invalid test instruments will be discarded (Efendi et al., 2024).

According to (Dachliyani, 2019), the test instrument must be reliable, reliable, steady, trustworthy, and not changeable, meaning that if the test instrument is used many times for the same subject at different times, it will get the same or relatively no other results. The test instrument developed has gone through a reliability test and was declared reliable so that the test instrument is consistent when used to measure various groups of students and is suitable for use (Sutami, 2020).

Discriminating power is the ability of an item of test instrument for learning outcomes to differentiate between students with high skills and those with low skills (Sutami, 2020). From the discriminating power analysis results, it can be concluded that the test instrument developed can differentiate between students who have mastered the material/learning achievement and those who have not.

According to (Sutami, 2020), difficulty analysis is needed to determine how difficult the instrument being tested is based on the test results carried out by students. The level of difficulty of test instruments helps create test instrument packages with an equivalent level of difficulty, for example, 25% difficult test instruments, 50% easy test instruments, and 35% medium test instruments; from the results of the level of difficulty analysis, it can be concluded that the test instruments developed can be used to meet the needs of test instruments with moderate and difficult levels of difficulty.

Distractor effectiveness analysis is used to determine the effectiveness of the answer choices answered by students as a basis for reviewing test instruments (Sutami, 2020). From the analysis results, it can be concluded that all test instrument items have alternative answers that function but with several revisions for several test instrument items to make them more effective.

The implication is to study and examine each item of test instruments to obtain quality questions by improving the quality of the test instruments through revision or discarding ineffective questions. Apart from that, item analysis also functions to find diagnostic information for students regarding the material achieved by the CP that has been implemented (Susanto et al., 2015). This statement is supported by (Fitrianawati, 2015); identification of the items of test instruments is carried out to obtain information, which is feedback to make improvements, improvements, and refinements to the items of test instruments so that they can measure what they want to measure. Zuriyanti (Fitrianawati, 2015) explains that the benefits of analyzing test instrument items are 1) determining which items of test instruments are not functioning correctly; 2) improving the quality of items of test instruments through discriminating power, level of difficulty, and distractor effectiveness; 3) increase validity and reliability; 4) make improvements to questions that are not relevant to the material being taught.

CONCLUSION

Based on the research results on 40 items of test instruments in the form of multiple choices, 30 test instruments were declared valid, and 10 test instruments were declared invalid. Thirty items of test instruments were declared reliable with a Cornbach's Alpha value of 0.805. The discriminating power of items of test instruments shows that 22 test instruments have "good" criteria and 8 items have "fairly good" criteria. The distractor effectiveness of each item of the test instrument is declared effective with each alternative answer chosen by the student. That proves that alternative answers can distract students from the correct answer. Test instruments that meet the criteria in CTT can be used to measure students' abilities in learning outcomes carried out at school. Moreover, test instruments with high-order thinking skills (HOTS) criteria can improve and familiarize students with higher-level thinking when solving tasks and problems while studying at school and working in the industry.

The limitation of this research is that the test instrument was only developed on simple financial management elements at MPLB Vocational Schools. The Future research is that test instruments can be developed for other elements in MPLB Vocational Schools so that Vocational Schools have a bank of HOTS test instruments that can be used jointly to measure students' abilities as well as analysis of test instrument development using Item Response Test (IRT)

ACKNOWLEDGMENTS

All funding for this research was borne by State University of Surabaya

REFERENCES

Ali, M. (2019). Analisis Butir Soal. *Journal of Chemical Information and Modeling*, *53*(9), 1689–1699. https://doi.org/10.13140/RG.2.2.26498.71360

Argina, A. W., Mitra, D., Ijabah, N., & Setiawan, R. (2017). Indonesian PISA Result: What Factors and What Should Be Fixed? *Proceedings Education and Language International Conference*, *1*(1), 69–79. https://jurnal.unissula.ac.id/index.php/ELIC/article/view/1212

Arikunto. (2013). Dasar-dasar Evaluasi Pendidikan. In *Bumi Aksara*.

Azwar, S. (2012). *Dasar-dasar Psikometri*. Pustaka Pelajar.

Bichi, A. A. (2015). Comparison of Classical Test Theory and Item Response Theory: A Review of Empirical Studies Test Items Development and Analysis Using Item Response Theory View project. *Australian Journal of Basic and Applied Sciences*, *9*(7), 549–556. https://doi.org/10.13140/RG.2.1.1561.5522

Cappelleri, J. C., Jason Lundy, J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical Therapeutics*, *36*(5), 648–662. https://doi.org/10.1016/j.clinthera.2014.04.006

Dachliyani, L. (2019). INSTRUMEN YANG SAHIH : Sebagai Alat Ukur Keberhasilan Suatu Evaluasi Program Diklat (evaluasi pembelajaran). *MEDIKA: Media Informasi Dan Komunikasi Diklat Kepustakawanan*, *5*(1), 57–65. https://ejournal.perpusnas.go.id/md/article/view/721/0

Desilva, D., Sakti, I., & Medriati, R. (2020). Pengembangan Instrumen Penilaian Hasil Belajar Fisika Berorientasi HOTS (Higher Order Thinking Skills) pada Materi Elastisitas dan Hukum Hooke. *Jurnal Kumparan Fisika*, *3*(1), 41–50. https://doi.org/10.33369/jkf.3.1.41-50

Efendi, T. N., Kartini, K., & Anggraini, R. D. (2024). Pengembangan Instrumen Tes Kemampuan Penalaran Matematis pada Materi Barisan dan Deret Kelas XI SMA/MA. *Jurnal Cendekia : Jurnal Pendidikan Matematika*, *8*(1), 811–826. https://doi.org/10.31004/cendekia.v8i1.2650

Fenanlampir, A., Batlolona, J. R., & Imelda, I. (2019). The struggle of Indonesian students in the context of TIMSS and Pisa has not ended. *International Journal of Civil Engineering and Technology*, *10*(2), 393–406.

Fitrianawati, M. (2015). Peran Analisis Butir Soal Guna Meningkatkan Kualitas Butir Soal, Kompetensi Guru Dan Hasil Belajar Siswa. *Prosiding Seminar Nasional Pendidikan PGSD UMS & HDPGSDI Wilayah Jawa*, *5*(3), 282–295.

Hutabarat, I. M. (2009). Analisis Butir Soal dengan Teori Tes Klasik dan Teori Respons Butir. *Pythagoras: Jurnal Pendidikan Matematika*, *5*(2), 1–13.

Kemendikbud. (2017). *Panduan Implementasi Kecakapan Abad 21 Kurikulum 2013 di Sekolah Menengan Atas*. Direktorat Pembinaan SMA Ditjen Pendidikan Dasar dan Menengah.

Lestari Puji Rahayu, A. S. N. D. (2018). Pengembangan Soal Matematika Hots (Higher Order Thinking Skills) Kelas X Berdasarkan Triple Theory. *Efektor*, *5*(2).

Magno. (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.

Muhson, A. (1979). Analisis Butir Soal Dengan Anbuso. *Jurnal Penelitian Pendidikan Guru Sekolah Dasar*, *6*(August), 128.

Pantiwati, Y. (2015). Strategi Pembelajaran, Self Assessment, Dan Metakognisi Dalam Pembelajaran Sains. *Prosiding Seminar Nasional Pendidikan Biologi 2015*, 677–685.

Pantiwati, Y. (2016). Hakekat Asesmen Autentik Dan Penerapannya Dalam Pembelajaran Biologi. *Jurnal Edukasi Matematika Dan Sains*, *1*(1), 18. https://doi.org/10.25273/jems.v1i1.773

Permana, I. (2018). *Implementation of Digital Assigments To Improve High Order Thinking Skills (HOTS) Ability of Senior High School Students In The Concept of Newton's Law*. *10*(2), 335–340. http://dx.doi.org/10.15408/es.v10i2.10236

Purbaningrum, K. A. (2017). Kemampuan Berpikir Tingkat Tinggi Siswa Smp Dalam Pemecahan Masalah Matematika Ditinjau Dari Gaya Belajar. *Jurnal Penelitian Dan Pembelajaran Matematika*, *10*(2), 40–49. https://doi.org/10.30870/jppm.v10i2.2029

Putri, R. A. H., & Pahlevi, T. (2021). Pengembangan instrumen penilaian berbasis hots berbantuan google form pada mata pelajaran kearsipan kelas x jurusan OTKP SMKN 2 Kediri. *Journal of*

Office Administration …, *1*(2), 138–152. https://ejournal.unesa.ac.id/index.php/joa/article/view/42123

Saccuzzo, K. &. (2009). *Psychological testing: Principles, applications, and issues* (7th ed.). wadsworth.

Serevina, V., Sari, Y. P., & Maynastiti, D. (2019). Developing high order thinking skills (HOTS) assessment instrument for fluid static at senior high school. *Journal of Physics: Conference Series*, *1185*(1). https://doi.org/10.1088/1742-6596/1185/1/012034

Sofyan, F. A. (2019). Implementasi Hots Pada Kurikulum 2013. *Inventa*, *3*(1), 1–9. https://doi.org/10.36456/inventa.3.1.a1803

Sudijono, A. (2013). *Pengantar Evaluasi Pendidikan*. Alfa Beta.

Sumaryanta. (2021). Teori Tes Klasik dan Teori Respon Butir: Konsep dan Contoh Penerapannya. In *Cetakan Pertama* (Vol. 15, Issue 2).

Supranoto, K. (2012). *Pengukuran dan Penilaian Pendidikan*. Graha Ilmu.

Susanto, H., Rinaldi, A., & Islam Negeri Raden Intan Lampung, U. (2015). Analisis Validity Reliability Level of difficulty dan Discriminating power pada Butir Soal Ujian Akhir Semester Ganjil Mata Pelajaran Matematika. *Jurnal Pendidikan Matematika*, *6*(2), 203–217.

Sutami. (2020). Pengembangan Instrumen Asesmen Higher Order Thinking Skills (HOTS) pada Mata Pelajaran Bahasa Indonesia SMA dan SMK. *Diglosia: Jurnal Kajian Bahasa, Sastra, Dan Pengajarannya*, *3*(1), 102–113. https://doi.org/10.30872/diglosia.v3i1.24

Thiagarajan, S. A. O. (1974). Instructional development for training teachers of exceptional children: A sourcebook. *Journal of School Psychology*, *14*(1), 75. https://doi.org/10.1016/0022-4405(76)90066-2

Umami, R., Rusdi, M., & Kamid, K. (2021). Pengembangan instrumen tes untuk mengukur higher order thinking skills (HOTS) berorientasi programme for international student asessment (PISA) pada siswa. *JP3M (Jurnal Penelitian Pendidikan Dan Pengajaran Matematika)*, *7*(1), 57–68. https://doi.org/10.37058/jp3m.v7i1.2069

Wang & Osterlind. (2013). *Classical Test Theory*. SensePublishers.

Wu, M., Tam, H. P., & Jen, T.-H. (2016). Educational Measurement for Applied Researchers. *Educational Measurement for Applied Researchers*. https://doi.org/10.1007/978-981-10-3302-5

INFORMATION ABOUT THE AUTHORS

Agnes Dwi Anggraeni : (Universitas Negeri Surabaya, Indonesia, agnesdwi.20006@mhs.unesa.ac.id)
Febrika Yogie Hermanto : (Universitas Negeri Surabaya, Indonesia, febrikahermanto@unesa.ac.id)