Jurnal Pendidikan Indonesia
Volume 12, Number 1, 2023 pp. 78-88
P-ISSN: 2303-288X E-ISSN : 2541-7207
Open Access: https://doi.org/10.23887/jpiundiksha.v12i1.52262

Empirical Quality of Final Exam Questions in a Learning Management System-Based Course

Hari Sugiharto Setyaedhi[1*], Mustaji[2], Citra Fitri[3]

[1,2,3]Faculty of Education, State University of Surabaya, Surabaya, Indonesia

ARTICLE INFO

Article history:
Received September 13, 2022
Revised September 15, 2022
Accepted January 12, 2023
Available online March 25, 2023

Kata Kunci:
Kualitas empirik, Ujian Akhir Semester, Pengembangan Media Grafis

Keywords:
Cmpirical quality, Final Semester Exam, Graphic Media Development

ABSTRAK

Masih banyak pendidik yang mengembangkan tes tanpa melalui tahapan analisis butir, sehingga belum diketahui kualitas butir soal. Hal ini berdampak pada keakuratan hasil pengukuran item. Penelitian ini bertujuan untuk menganalisis kualitas empirik butir soal ujian akhir mata kuliah pengembangan media grafis. Metode penelitian yang digunakan adalah metode deskriptif dengan pendekatan kuantitatif. Teknik pengumpulan data adalah analisis lembar jawaban tugas siswa pada ujian akhir. Populasi terdiri dari siswa dengan jumlah 90 siswa. Data dianalisis dengan menggunakan aplikasi Microsoft Excel 2010 untuk mengetahui tingkat kesukaran, daya pembeda, keefektifan distraktor, validitas, dan reliabilitas. Hasil penelitian kualitas empiris dari total 25 soal pilihan ganda untuk indeks kesukaran soal. Kemudian untuk keefektifan distraktor terdapat 15 (60%) item efektif dan 10 (40%) item tidak efektif. Sedangkan untuk validitas item terdapat 24 (96%) item yang masuk kategori valid dan hanya 1 (4%) item yang masuk kategori tidak valid. Reliabilitas tes dengan menggunakan Cronbach's alpha berada pada kategori sangat tinggi yaitu 0,812. Dengan demikian dapat disimpulkan bahwa UN semester mata kuliah pengembangan media grafis layak digunakan sebagai alat untuk mengukur kompetensi mahasiswa.

ABSTRACT

There are still many educators who develop tests without going through the stages of item analysis, so the quality of the test items is unknown. This has an impact on the accuracy of the item measurement results. This study aims to analyze the empirical quality of the question items in the final examination of the graphic media development course. The research method used was a descriptive method with a quantitative approach. The data collection technique was the analysis of the answer sheets of student work on the final examinations. The population was made up of students with a total of 90 students. The data were analyzed using the Microsoft Excel 2010 application to determine the level of difficulty, discriminating power, distractor effectiveness, validity, and reliability. The empirical quality research results from a total of 25 multiple-choice items for the item difficulty index. Then, for the effectiveness of the distractor, there were 15 (60%) effective items and 10 (40%) ineffective items. Meanwhile, for the validity of the items, there were 24 (96%) items in the valid category and only 1 (4%) item in the invalid category. The reliability of the test using Cronbach's alpha was in the very high category, namely 0.812. Thus, it can be concluded that the semester final examination for graphic media development courses is appropriate to be used as a tool to measure student competency.

1. INTRODUCTION

To achieve national education objectives (sisdiknas), all components of the national education system must be viewed as interconnected units. Sisdiknas is expected to have success without discrimination (Khunaifi & Matlani, 2019; Rahman et al., 2021). National education goals contain the values to be realized in educational processes or activities. The goal of national education is to develop capabilities and form dignified national character and civilization in the context of educating the nation's life, with the goal of developing students' potential to become human beings who believe and fear God Almighty, have noble character, are healthy, knowledgeable, capable, creative, independent, and democratic and responsible citizens (Rukiyati, 2019; Sujana, 2019). The general mission of Law No. 12 of 2012 is also to prepare students to become members of society who have academic and/or professional

*Corresponding author
E-mail addresses: harisugii89@gmail.com (Hari Sugiharto Setyaedhi)

abilities and can apply, develop, and/or enrich the treasures of science, technology, and/or the arts and strive for their use to improve people's lives. and contribute to national culture (Nursanjaya, 2019). The quality of education in tertiary institutions can be achieved through the processes that occur in planning and presenting lecture material, evaluating processes, products, and elements involved in efforts to meet the needs of everyone involved, especially students and the world of work. Evaluation is an activity of controlling, guaranteeing, and determining the quality of education for various components of education in every path, level, and quality of education in learning that is carried out through interaction activities between students and lecturers in a learning environment (Holiah, 2022; Maisah et al., 2020; Pangalila, 2017). Learning evaluation is carried out in order to monitor learning outcomes. This sisdiknas law can be used to evaluate student learning outcomes conducted by lecturers at tertiary institutions. (Idrus, 2019; Warju et al., 2020). Lecturers must at least master four competencies well, namely: 1) master substance; 2) master methodology; 3) master evaluation techniques; and 4) understand, live, and practice moral values and the professional code of ethics (Anetha & Hasriyanti, 2019; Riswanda & Burhan, 2022).

Measurement, assessment, and evaluation are carried out in stages, meaning that we must first carry out measurement activities, namely comparing observations with criteria, before the measurement results are interpreted in the assessment process. Furthermore, the evaluation applies the value so that a decision can be taken (Tarmizi et al., 2021; Zainal, 2020). The target of evaluation in the world of education can be in the form of learning outcomes. Learning outcomes are a series of student learning processes that have taken place within a certain period of time (Gunawan et al., 2018; Magdalena et al., 2020). In analyzing and measuring student learning outcomes, a process of assessing learning outcomes is needed, which is carried out systematically and continuously using an assessment instrument, namely a test. The test is a procedure that needs to be taken within the framework of measurement and assessment in the field of education (Magdalena et al., 2021; Syachlani & Setyorini, 2021). The test is a planned measurement that is used to provide opportunities for students to show their achievement results related to predetermined goals. The test as a measuring tool is specifically designed with learning objectives and must be prepared as well as possible in accordance with predetermined rules. In the evaluation process, a test of good quality is needed to determine the quality of the data produced (Kurniawati, 2019; Pangesti et al., 2020).

Item difficulty level, item discrimination, item distractor effectiveness, item validity, and test reliability are all examined empirically. The validity of the items and the reliability of the test include the difficulty level of the items, their discrimination, and the effectiveness of the item distractors (Iskandar & Rizal, 2018; Srika Ningsih Pasi; Yusrizal, 2018; Widayanti et al., 2021). Item analysis is a systematic process that will provide very specific information on the items compile. Test item analysis is one of the activities that needs to be carried out in order to improve the quality of the questions that have been prepared and aims to identify good, bad, and very bad questions (Farida & Musyarofah, 2021; Nur & Palobo, 2018). Item analysis needs to be carried out by each lecturer in each subject that will be tested. Item analysis can be done qualitatively or quantitatively. A qualitative analysis of the items or a theoretical analysis is carried out before the items are tested and empirically analyzed.

Item analysis needs to be done to test the quality of each item, and a set of questions in various aspects is needed. However, because developing quality items is not easy, most lecturers evaluating students' cognitive domains only carry out qualitative item analysis, while empirical analysis is still rarely carried out (Elviana & Murdiono, 2017; Erawati, 2018). The stages of developing test instruments are: 1) compiling test specifications; 2) making or writing tests; and 3) discussing the tests in the stages of designing and developing test instruments. 4); trial tests; 5); test analysis; 6); test fixing; 7); test assembly; 8); test execution; and 9); interpretation of test results (Fitrianawati, 2017; Ndiung & Jediut, 2020; Nurul R.A. et al., 2021). One of the steps in developing a test instrument is to analyze the items. This is done solely for the purpose of obtaining a good or high-quality test instrument. A good test instrument has the following characteristics: validity, reliability, relevance, representativeness, practicality, discriminatoryness, specificity, and proportionality (D. D. Kurniawan, 2014; Zainal Arifin, 2019). A good test is valid, dependable, objective, practicable, cost-effective, representative, varied, discriminatory, and meaningful. A good test has the following conditions: the test must be valid, the test must be reliable, the test must be objective, and the test must be discriminatory (Arikunto, 2018; Suharman, 2018). A good test must meet the following criteria: a) reliability, b) validity, c) objectivity, d) discriminatoryness, e) comprehensiveness, and f) ease of use. This is in accordance with the thinking of measurement experts, who define the main criteria for measuring instruments (instruments) used in making measurements, which are psychological in nature, namely validity and reliability (Purniasari et al., 2021; D. Widiyanto & Istiqomah, 2020). The results of observations in the Curriculum and Educational Technology Department found the following problems. So far, UAS in the graphic media development course uses an essay test that contains many weaknesses, such as: a) scoring is often

influenced by the subjectivity of the assessor; b) it takes a long time to correct answers; and c) the scope of the material being tested is very limited. Lecturers develop tests without going through item analysis; this has an impact on the quality of the items. Previous research said that there are still many educators who develop tests without going through the stages of item analysis, so the quality of the test items is unknown (Faridah, 2021). This has an impact on the accuracy of the item measurement results. Research conducted by other study said that most students scored below 60 because the quality of the questions used in the UAS was unknown, so it was necessary to analyze the items (Utami et al., 2020). Some of the educational technology lecturers do not understand how to analyze the items. This is in accordance with research said that the knowledge and skills of educators in analyzing the items is still low (Sumiati et al., 2018). Most educators tend to ask questions without first measuring whether students have understood the material to be tested, so that it can be ascertained that student competency cannot be measured precisely. Currently, a lot of education is more concerned with results without knowing the process that the students themselves go through (R. Y. Kurniawan et al., 2017; Zahiroh & Ritonga, 2021).

The solution to overcome this problem Lecturers must use multiple-choice UAS questions (Multiple Choice Test). Multiple-choice tests are widely used in education. Multiple-choice questions have several advantages, namely: 1) they can measure various cognitive levels, 2) Scoring multiple-choice tests is easy, fast, and objective; they cover a broad scope of material. 3) Multiple-choice tests are appropriate for exams with a large number of participants, and the results can be announced immediately (Destiniar et al., 2018; Sanusi & Aziez, 2021). Multiple choice tests make it easier for lecturers to analyze the difficulty level of the items before the questions are tested so that the validity of the items can be determined. Thus, the test will produce accurate and reliable scores (Agustiana et al., 2019; Ariyanti & Bhakti, 2020; Tarmizi et al., 2021). Analysis of the items aims to determine the quality of the items. The exam questions that will be used must be tested for their feasibility to determine the quality of each item. In general, a good test instrument must pay attention to the level of difficulty, item discrimination, the effectiveness of the item distractor, the validity of the item and the reliability of the test (Nengsi & Efrina, 2019; Supiyansyah et al., 2017). A good test must be valid, objectively reliable, practicable and practical. Therefore this study aims to analyze the empirical quality of the question items in the final examination of the graphic media development course.

2. METHOD

The type of data used in this study is quantitative data expressed in numbers. This research was conducted at the Department of Curriculum and Educational Technology, Faculty of Education, State University of Surabaya. The data collection method used in this study was documentation in the form of responses or answers to UAS tests in the graphic media development course for 90 Curriculum and Educational Technology students. The UAS script for the course on graphic media development is in the form of multiple choice using five options with a total of 25 items. The item analysis was carried out in a quantitative and descriptive manner. Descriptive analysis is the analysis used to analyze data by describing the data that has been collected (Anetha & Hasriyanti, 2019). Data from student answers are summarized, and then an empirical quality analysis is carried out. The data analysis technique used in this study is Microsoft Excel. This study will analyze the level of difficulty of the items, the different powers of the items, the effectiveness of the distractor, the validity of the items, and their reliability. The difficulty level of the item can be seen from the size of the number, which is expressed in terms of the item difficulty index number (difficulty index), which is generally denoted by the letter P (proportion). The item difficulty index of the items ranges from 0.00 to 1.00. The item diffivulity index is show in Table 1.

Table 1. Item Difficulty Index (P)

P	Interpretasi
< 30	Too Difficult
0,30 – 0,70	Enough (Moderate)
> 70	Too easy

The item discrimination index or item discrimination index is the ability of the item to distinguish groups with high ability (upper group) and groups with low ability (lower group) (Bagiyono, 2017; Supriyati & Dudung, 2019). The discrimination index (D) ranges from 0.00 to 1.00. In this index, it is possible that the value of the discrimination index is negative. The criteria used to determine the discrimination index of these questions are presented in Table 2.

Table 2. Discrimination Index Criteria (D)

Discrimination index	Classification	Interpretation
< from 0,20	Bad	Items are considered not to have good discriminatory power
0,21 – 0,40	Currently	The items have sufficient (moderate) discriminatory power
0,40 – 0,70	Well	These items have good discriminatory power
0,70 – 1,00	Very well	These items have excellent discriminatory power
Negative sign	-	These items have very poor discriminatory power

The reliability coefficient has a range from 0.00 to 1.00. Obtaining an Alpha score in a program with a reliability classification of 0.00 – 0.20 (very low), 0.21 – 0.40 (low), 0.71 – 0.90 (high), and 0.91 – 1.00 (very high) (Nuryanti et al., 2018; Sanaky, 2021). The test instrument is declared reliable if the reliability coefficient has a minimum value of 0.6 (Fatimah & Alfath, 2019; Syaifudin, 2020; Zahiroh & Ritonga, 2021). Cronbach's alpha has three categories, namely $\alpha < 0.7$ less convincing reliability, $\alpha > 0.7$ reliability in the good category, and $\alpha > 0.8$ reliability in the special category. The reliability category can be seen in Table 3.

Table 3. Reliability Categorization

Test Reliability Coefficient	Category
$\alpha < 0,7$	Less Reliable Reliability
$\alpha > 0,7$	Good Category Reliability
$\alpha > 0,8$	Special Reliability

3. RESULT AND DISCUSSION

Result

This research was conducted with the aim of knowing empirical quality, namely the level of difficulty of the items, item discrimination, the effectiveness of the item distractors, the validity of the items, and the reliability of the UAS test questions for the graphic media development course on 25 multiple-choice items using 5 options. The responses from 90 students were then analyzed so that the empirical quality of the UAS in the graphic media development course was known. The results of the empirical quality of the item analysis are as follows.

Item difficulty Index

Analysis of the level of difficulty of the items means analyzing the items in order to obtain items that fall into the categories of easy, medium and difficult. The results of the analysis of the Difficulty Index (P) for UAS in the development of graphic media can be seen in the Table 4.

Table 4. Item Difficulty Level Index Categorization

Index P	Category	Items	Amount	Percentage
< 0,30	Too difficult	24	1	4%
0,30 – 0,70	Medium	5, 9, 16, 19, 20, 21, 22	7	28%
> 0,70	Too Easy	1, 2, 3, 4, 6, 7, 8, 10, 11, 12, 13, 14, 15, 17, 18, 23, 25	17	68%

The difficulty index (P) of the items on the 25 UAS questions in the graphic media development course showed that 1 (4%) item was in the "too difficult" category, namely item number 24, and then 7 (28%) items were in the "medium" category, namely item numbers 5, 9, 16, 19, 20, 21, and 22. As for the items that were "too easy," there were 17 (68%) items, namely item numbers: 1, 2, 3, 4, 6, 7, 8, 10, 11, 12, 13, 14, 15, 17, 18, 23, and 25.

Item Discrimination Index (D)

The results of the analysis of the item discrimination index (D) for the Final Semester Examination (UAS) for the development of graphic media can be seen in the Table 5.

Table 5. Discrimination Index (D) and Classification

Index D	Category	Items	Amount	Percentage
< dari 0.20	Bad	2, 4, 7, 13, 16, 20, 24	7	28%
0.21 – 0.40	Currently	1, 3, 6, 8, 9, 10, 11,12, 14, 15, 18, 21, 23	13	52%
0.40 – 0.70	Well	5, 17, 19, 22, 25	5	20%
0.70 – 1.00	Very well			
negatif	Ugly			

Base on Table 5, the item discrimination index (D) for the 25 UAS questions in the graphic media development course shows that there are 7 items in the bad category, namely item numbers: 2, 4, 7, 13, 16, 20, and 24 (28%), while D is in the moderate category. There are 13 items, namely item numbers: 1, 3, 6, 8, 9, 10, 11, 12, 14, 15, 18, 21, and 23 (52%), while D is in the good accepted category, there are 5 items, namely item number: 5, 17, 19, 22, and 25 (20%).

The Effectiveness of the item distractor

Distractors or options outside the correct answer key are said to be effective if chosen by at least 5% of all students. Items are said to be effective if at least 5% of students are selected. The effectiveness of the distractor of the Semester Final Examination for the graphic media development course can be seen in Table 6.

Table 6. Distractors Categorization

Nilai	Category	Items	Amount	Percentage
≥ 5%	Effective	5, 6, 8, 11, 12, 14, 15, 17, 18, 20, 21, 22, 23, 24, 25	15	60%
≤ 5%	Ineffective	1, 2, 3, 4, 7, 9, 10, 13, 16, 19	10	40%
0.000	Not good/reject	-		

Base on Table 6, the effectiveness of the distractor on the 25 UAS questions in the graphic design development course shows that there are 15 items of the distractor that are effective, namely items numbers 5, 6, 8, 11, 12, 14, 15, 17, 18, 20, 21, 22, 23, 24, and 25 (60%), while the items that distracted were not effective, namely items numbers 1, 2, 3, 4, 7, 9, 10, 13, 16, and 19 (40%).

Item Validity

Validity generally refers to the fact that a measuring instrument has precision or accuracy in measuring what is to be measured. Test instruments can be used if the instrument is capable of producing the same results to evaluate a measurement. The higher the validity value indicates the more accurate a measuring instrument is for measuring data. This validity test is important so that the questions given produce valid data. In this study, the product moment correlation formula (rxy) was used to test the results of the item validity analysis, and the correlation index rxy was then consulted with r tables at a significance level of 5% according to db (degrees of freedom), namely the number of students (N-2) or 90 - 2 and obtained an r table with a sig 5% of 0.207. Items are said to be valid if their rcount is greater than 0.207 and invalid if their rcount is less than 0.207. The validity of the test items from the UAS graphic media development course can be seen in the Table 7.

Table 7. Price Categorization r

Index Validity	Items	Amount	Percentage
Valid Items r_{pbis} > 0,207	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 19, 20, 21, 22, 23, 24, 25	24	96%
Invalid Items r_{pbis} < 0,207	16	1	4%

Base on Table 7, the validity of the 25 UAS items in the graphic design development course contained 24 items which were declared valid or 96%, while only 1 item was declared invalid, namely number 16 or 4%.

Reliabilitas Tes

The reliability results using Cronbach's Alpha in the Semester Final Examination for the graphic media development course can be seen in the Table 8.

Table 8. Reliability Results

Cronbach's Alpha	N of Items
0.812	25

Base on Table 8, the analysis of the items obtained from the reliability test yielded a reliability of 0.812, which means exceptional. Thus, the test instrument is very feasible to be used as UAS material for graphic media development courses.

Discussion

The quality of the items is seen from the empirical quality of the items, which consists of the level of difficulty of the items, item discrimination, the effectiveness of the item distractors, the validity of the items, and the reliability of the test. The following is a discussion of each of the characteristics of the item assessment.

Items difficulty

The results showed that, for the difficulty level of the items, there were 1 (4%) that were in the "too difficult" category and had to be revised, 7 (28%) that were good because they were in the "medium" category, and 17 (68%) that were in the "too easy" category. and must be revised. Based on the data above, the items with easy and difficult difficulty levels need to be examined and revised again (Herkusumo, 2011; Widayanti et al., 2021). Good items are questions that are neither too easy nor too difficult; questions that are too easy do not motivate or stimulate students to solve problems, whereas questions that are too difficult frustrate, hopeless, and leave students with no desire to try again (Arikunto, 2018; Widyawati, 2017). Medium-category items are good questions, which means that students with high ability can work on the items correctly and students with low ability will have difficulty answering the items. Following the analysis of the items for their level of difficulty, the following follow-up can be performed: 1) Items in the medium category or good questions are used as a question bank; 2) items with categories that are too difficult can be discarded or dropped. Items 2, 4, 7, and 13 are in the bad category because they are too easy. There are several possibilities why the item is in the category of "too easy," namely: a) The question on the item using the answer option is too easy for students to guess, or the question points to one of the answer keys. b) Most of the students answered the item correctly, meaning that most of the students had understood the material being asked. c) the distractor on the item did not work at all, so that almost all students could answer; d) the item or answer was leaked before the test was tested on students. Meanwhile, items that are too difficult may be caused by students not learning or not learning optimally. One of the objectives of item analysis is to determine which questions are flawed and not functioning (Fitrianawati, 2017; Utami et al., 2020).

Item discrimination

Discrimination in items in the bad category was 7 items (28%), discrimination in items in the moderate category was 13 items (52%), and discrimination in items in the good category was 5 items (20%). Discrimination of items in the moderate and good categories, which has been included in the question bank and can be reused for future tests (Arikunto, 2018; Vincent & Shanmugam, 2020). Based on the results of the discrimination analysis of the item items, it can be concluded that the UAS item items for the graphic design development course are able to distinguish high- and low-ability students. Previous study state in principle, the discriminating power of the items reflects the differences in the answers to the items between groups of students with high abilities and those with low abilities (Qurrota et al., 2022). Smart students have a great chance to answer the questions correctly compared to those who are less intelligent. Item numbers 2, 7, 13, 16, 20, and 24 are items with discrimination; items in the bad category, on the contrary, are discarded because they cannot distinguish between students who are good at them and students who are not good at them (O. A. & E. R. I., 2016; Widayanti et al., 2021). If an item cannot distinguish between the two student abilities, then the item can be suspected of "probability" as follows: 1) The answer key to the item is incorrect; 2) The item has two or more correct answer keys; 3) Competency What was measured was unclear; 4) the distractor did not work; 5) the material asked was too difficult, so many test takers guessed; 6) most of the test takers who understood the material thought there were errors in the items.

Effectiveness of Distractors

The effectiveness of item distractors in the "effective" category is 60% (15 items), and the "ineffective" category is 40% (10 items). A good distractor is one that has stimulating or seducing power, causing students, particularly those with weak, low, or less intelligent thinking abilities, to hesitate to

answer correctly. Items with good distractors will be chosen evenly by students who answer incorrectly. On the other hand, items that are not well distracted will be chosen unevenly by students. In line with previous researcher statement that ideally, the distractor should be chosen only by incompetent or incapable subjects, while none of the capable subjects choose the distractor (Arbiatin & Mulabbiyah, 2020). As with the answer key, of course, in reality, there is still a chance that a capable subject chooses the wrong distractor. If the proportion remains smaller than the proportion of distractor voters from the incompetent subject group, then the distractor can still be considered effective.

Items 5, 6, 8, 11, 12, 14, 15, 17, 18, 20, 21, 22, 23, 24, and 25 are items with effective distractors and are good items. The distractor is said to be good if all distractions are functioning; the distractor is said to be in a good category if it has 1 distraction that does not work; the category is bad if it has 2 distractions that do not work; the category is poor if it has 3 distractions that do not work; and the category is very bad if there are 4 distractions. does not work. (Arbiatin & Mulabbiyah, 2020). The more students choose the distractor, the more it has carried out its function properly (J. Widiyanto, 2018). Item numbers 1, 2, 3, 4, 7, 9, 10, 13, 16, and 19 are items with distractors that are not effective or don't work properly. The distraction has no appeal for students who do not understand the material. The poor quality of the distractor is caused by the distractor being too conspicuous or misleading (Prawiki & Helendra, 2022). In order to function properly, the distractor must be made as similar as possible to the answer key. A bad distractor is a distractor that is not chosen at all by students because it looks too misleading. As a result, it is difficult for the question creator to create a distraction so that the answer is difficult for students to guess correctly.

Items Validity

The results of the analysis of the items indicate that the validity of the items in the "valid" category is 96% (24 items). This means that most of the items can distinguish students who have achieved learning objectives from those who have not. Thus, UAS in the graphic media development course has been able to measure what should be measured, not anything else (D. D. Kurniawan, 2014; Nurul R.A. et al., 2021). Validity is affected. 1) insufficient time to work on the questions; 2) cheating in the execution of the test; 3) inconsistent scoring; 4) test takers did not follow the directions given in the test; 5) there was a jockey. Items can be valid if the construction is good and includes material that represents the target measure (Anetha & Hasriyanti, 2019; Prawiki & Helendra, 2022). Previous study said that one of the factors that influences validity is the answer factor from students by guessing answers (guessing) (Ardhani, 2020). So it can be concluded that UAS items for the graphic media development course are appropriate for measuring what should be measured. Item number 16 is said to be invalid. There are three factors that affect the validity of the test results: the instrument used for the test, the administration and scoring factors, and the factor of student answers (Iskandar & Rizal, 2018; Zainal Arifin, 2019). Previous study said obtaining validity based on the student's ability to answer per test item will then be calculated by the total number of correct student answers, and a validity value will be obtained for test item number 1 (Maulida & Hamama, 2021). Student ability greatly affects the validity of the test items, and besides that, the correct total of all test items will affect the validity value. If there are many students who can answer the test items correctly, then the validity value of the test items will be high, and vice versa, if only a few students are able to answer the test items, then the validity value will decrease as well as the validity.

Test Reliability

The results of the analysis of UAS items in the graphic media development course have a Cronbach Alpha value of 0.812, which is classified as "very high," meaning that the test has very high consistency. The consistency in question includes the accuracy of the measurement results and the stability of the measurement results. The high value of reliability is very closely related to the validity of the items. The relationship between validity and reliability concerns the accuracy of the test in measuring the symptoms to be measured, while reliability refers to the extent to which a measurement can be trusted or consistent (Utami et al., 2020; Yusup, 2018). Tests that are valid and measure what should be measured will definitely show consistent or reliable results, but consistent measurement terms cannot show support for validity. However, validity is more important than reliability. Reliability has an influence on validity; therefore, a valid measuring instrument is always reliable, but a reliable measuring instrument is not necessarily valid (Arikunto, 2018; Sugiyono, 2019). It is important for a test to have validity and reliability requirements. The test may be reliable but not valid. On the other hand, a valid test is definitely reliable. The items used to measure student abilities need to pay attention to their quality, including items that must be valid and reliable. Besides that, items are said to be good if they are not too easy or too difficult, and they must also be able to distinguish between students who are smart and those

who are not. Previous study state clever, and the effectiveness of the distraction works well (Friatma & Anhar, 2019).

4. CONCLUSION

Considering the empirical quality of the UAS test items in the graphic media development course, which consists of the level of difficulty of the items, item discrimination, the effectiveness of item distraction, the validity of the items, and the reliability of the test, it can be concluded that the final semester exam items are feasible to test. The item difficulty index is mostly in the moderate category, so it is feasible to be tested. The discrimination index of the items was able to distinguish students with high and low abilities in the majority of the items. The effectiveness of the item distractors for some items still requires revision. The validity of the items is very good, meaning that the test has measured what it should measure. UAS reliability in graphic design development courses using Alpha Cronbach has very high consistency. Items that are considered good are maintained, while items in the bad category need to be revised or discarded. Basically, the UAS question items in the graphic media development course are very good; this is proven by the fact that most of the test items are valid and have very high reliability. Thus, the UAS graphic design development course deserves to be called UAS.

5. REFERENCES

Agustiana, M., Mayrita, H., & Muchti, A. (2019). Analisis Butir Soal Ulangan Akhir Semester Mata Pelajaran Bahasa Indonesia Kelas Xi. *Jurnal Ilmiah Bina Edukasi*, *11*(01), 26–35. https://doi.org/10.33557/jedukasi.v11i01.203.

Anetha, & Hasriyanti. (2019). Analisis Butir Soal Semester Ganjil Mata Pelajaran Matematika pada Sekolah Menengah Pertama. *Jurnal Pengukuran Psikologi Dan Pendidikan Indonesia (JP3I)*, *8*(1), 57–68. https://doi.org/10.15408/jp3i.v8i1.13068.

Arbiatin, E., & Mulabbiyah, M. (2020). Analisis Kelayakan Butir Soal Tes Penilaian Akhir Semester Mata Pelajaran Matematika Kelas Vi Di Sdn 19 Ampenan Tahun Pelajaran 2019/2020. *El Midad*, *12*(2), 146–171. https://doi.org/10.20414/elmidad.v12i2.2627.

Ardhani, Y. (2020). Pelajaran Teknologi Dasar Otomotif Kelas X Teknik Kendaraan Ringan Otomotif di SMK Muhammadiyah Gamping Periode 2018/2019. *Jurnal Pendidikan Vokasi Otomotif*, *3*(1), 85–94. https://doi.org/10.21831/jpvo.v3i1.34917.

Arikunto, S. (2018). *Dasar - Dasar Evaluasi Pendidikan*. Bumi Aksara.

Ariyanti, E., & Bhakti, Y. B. (2020). Perbandingan Bentuk Tes Pilihan Ganda dan Teknik Penskoran Terhadap Reliabilitas Tes Mata Pelajaran Kimia. *Titian Ilmu: Jurnal Ilmiah Multi Sciences*, *12*(2), 66–76. https://doi.org/10.30599/jti.v12i2.627.

Bagiyono. (2017). Analisis Tingkat Kesukaran dan Daya Pembeda Sial Ujian Pelatihan Radiografi Tingkat 1. *Widyanuklida*, *16*(1), 1–12. http://jurnal.batan.go.id/index.php/widyanuklida/article/view/4068.

Destiniar, D., Octaria, D., & Mulbasari, A. S. (2018). Analisis Butir Soal Pilihan Ganda Dengan Aplikasi Klasika. *J-ABDIPAMAS (Jurnal Pengabdian Kepada Masyarakat)*, *2*(1), 21. https://doi.org/10.30734/j-abdipamas.v2i1.180.

Elviana, P. S. O., & Murdiono, M. (2017). Pengaruh metode sosiodrama terhadap hasil belajar dan sikap tanggung jawab dalam pembelajaran PKn. *Jurnal Civics: Media Kajian Kewarganegaraan*, *14*(1), 33–50. https://doi.org/10.21831/civics.v14i1.14560.

Erawati, N. K. (2018). Analisis Tes Penilaian Pencapaian Kompetensi Pada Mahasiswa Kebidanan. *Jurnal Penjakora*, *5*(2), 111–120. https://doi.org/10.23887/penjakora.v5i2.17287.

Farida, & Musyarofah, A. (2021). Validitas dan Reliabilitas dalam Analisis Butir Soal. *Al-Mu'arrib: Jurnal Pendidikan Bahasa Arab*, *I*(1), 34–44. https://doi.org/10.32923/al-muarrib.v1i1.2100.

Faridah, A. (2021). Karakteristik Butir Soal Ujian Akhir Semester Mata Pelajaran Sejarah. *Ekspose: Jurnal Penelitian Hukum Dan Pendidikan*, *20*(2), 1281–1288. https://doi.org/10.30863/ekspose.v20i2.1819.

Fatimah, L., & Alfath, K. (2019). Analisis Kesukaran Soal, Daya Pembeda dan Fungsi Distraktor. *Jurnal Komunikasi Dan Pendidikan Islam*, 37–64. https://doi.org/10.36668/jal.v8i2.115.

Fitrianawati, M. (2017). Peran Analisis Butir Soal Guna Meningkatkan Kualitas Butir Soal, Kompetensi Guru Dan Hasil Belajar Peserta Didik. *JPT : Jurnal Pendidikan Tematik*, *2*(3), 316–322. https://publikasiilmiah.ums.ac.id/xmlui/handle/11617/9117.

Friatma, & Anhar. (2019). Analysis of validity, reliability, discrimination, difficulty and distraction effectiveness in learning assessment. *Journal of Physics: Conference Series*, *1387*(1). https://doi.org/10.1088/1742-6596/1387/1/012063.

Gunawan, Kustiani, L., & Hariani, L. S. (2018). Faktor Yang Mempengaruhi Hasil Belajar Siswa. *Mimbar Ilmu*, *26*(2), 193. https://doi.org/10.23887/mi.v26i2.35688.

Herkusumo, A. P. (2011). Penyetaraan (Equating) Ujian Akhir Sekolah Berstandar Nasional (UASBN) dengan Teori Tes Klasik. *Jurnal Pendidikan Dan Kebudayaan*, *17*(4), 455. https://doi.org/10.24832/jpnk.v17i4.41.

Holiah, I. (2022). Penguatan Kompetensi Guru Melalui Pengembangan Keprofesian Berkelanjutan. *Eduvis : Jurnal Manajemen Pendidikan Islam*, *7*(1), 97–106. https://www.neliti.com/publications/376667/penguatan-kompetensi-guru-melalui-pengembangan-keprofesian-berkelanjutan.

Idrus. (2019). Evaluasi Dalam Proses Pembelajaran. *Evaluasi Dalam Proses Pembelajaran*, *2*, 920–935. https://garuda.kemdikbud.go.id/documents/detail/1655265.

Iskandar, A., & Rizal, M. (2018). Analisis kualitas soal di perguruan tinggi berbasis aplikasi TAP. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *22*(1), 12–23. https://doi.org/10.21831/pep.v22i1.15609.

Khunaifi, A. Y., & Matlani, M. (2019). Analisis Kritis Undang-Undang Sisdiknas Nomor 20 Tahun 2003. *Jurnal Ilmiah Iqra'*, *13*(2), 81. https://doi.org/10.30984/jii.v13i2.972.

Kurniawan, D. D. (2014). Analisis Butir Soal Ujian Akhir Semester Matematika Dengan Teori Respon Butir. *Jurnal Riset Dan Konseptual*, *4*(20), 215–224. https://doi.org/10.28926/briliant.v3i.

Kurniawan, R. Y., Fiky Prakoso, A., Hakim, L., Mustika Dewi, R., & Widayanti, I. (2017). Pemberian Pelatihan Analisis Butir Soal Bagi Guru di Kabupaten Jombang : Efektif? *Jurnal Pemberdayaan Masyarakat Madani (JPMM)*, *1*(2), 179–193. https://doi.org/10.21009/jpmm.001.2.03

Kurniawati, A. (2019). Analisis Hasil Tes Evaluasi Pendidikan Pada Mahasiswa Ditinjau Dari Perbedaan Gender. *Jurnal Ilmiah DIDAKTIKA*, *19*(1), 89–106. https://doi.org/10.22373/jid.v19i1.4196.

Magdalena, I., Fauziah, S. N., Faziah, S. N., & Nupus, F. S. (2021). Analisis Validitas, Reliabilitas, Tingkat Kesulitan dan Daya Beda Butir Soal Ujian Akhir Semester Tema 7 Kelas III SDN Karet 1 Sepatan. *BINTANG : Jurnal Pendidikan Dan Sains*, *3*(2), 198–214. https://doi.org/10.36088/bintang.v3i2.1291.

Magdalena, I., Maula, N. H., Amelia, S. A., & Ismawati, A. (2020). Evaluasi Penerapan Pembelajaran K13 di Sekolah Dasar Dharmawati Arief Tangerang. *MANAZHIM*, *2*(1). https://doi.org/10.36088/manazhim.v2i1.596.

Maisah, Fauzi, H., Aprianto, I., Amiruddin, A., & Zulqarnain. (2020). *Strategi Pengembangan Mutu Perguruan Tinggi*. *1*(5), 416–424. https://doi.org/10.31933/JIMT.

Maulida, & Hamama, S. F. (2021). Pengembangan Instrumen Tes Tipe Pilihan Ganda Dalam Evaluasi Hasil Belajar Siswa Pada Konsep Sel Tingkat Sekolah Menengah Atas. *Jurnal Dedikasi Pendidikan*, *5*(1), 171–178. https://doi.org/10.30601/dedikasi.v5i1.1498.

Ndiung, S., & Jediut, M. (2020). Pengembangan instrumen tes hasil belajar matematika peserta didik sekolah dasar berorientasi pada berpikir tingkat tinggi. *Premiere Educandum: Jurnal Pendidikan Dasar Dan Pembelajaran Volume*, *10*(June), 94–111. https://doi.org/10.25273/pe.v10i1.6274.

Nengsi, A. R., & Efrina, G. (2019). Optimasi validitas dan reliabilitas tes pilihan ganda buatan guru mata pelajaran ips sd. *Journal Innovation in Islamic Education: Challenges and Readiness in Society 5.0, 4th International Conference on Education*, 43–48. https://ojs.iainbatusangkar.ac.id/ojs/index.php/proceedings/article/view/2138.

Nur, A. S., & Palobo, M. (2018). Pelatihan Analisis Butir Soal Berbasis Komputerisasi Pada Guru SD. *MATAPPA: Jurnal Pengabdian Kepada Masyarakat*, *1*(1), 5–11. https://doi.org/10.31100/matappa.v1i1.79.

Nursanjaya. (2019). Eksistensi Pendidikan Tinggi di Indonesia : Idealisme Atau Bisnis? *Negotium: Jurnal Ilmu Administrasi Bisnis*, *2*(1), 21–33. https://doi.org/10.29103/njiab.v2i1.3026.

Nurul R.A., Abdul, H., & Nurul, M. (2021). Analisis Butir Soal Matematika Pada Siswa Sekolah Dasar. *Pinisi Journal of Education*, *1*(1). https://doi.org/10.37058/jarme.v3i1.2501.

Nuryanti, L., Zubaidah, S., & Diantoro, M. (2018). Analisis Kemampuan Berpikir Kritis Siswa SMP. *Jurnal Pendidikan: Teori, Penelitian, Dan Pengembangan*, *3*(2), 155–158. https://doi.org/10.17977/jptpp.v3i2.10490.

O. A., A., & E. R. I., A. (2016). Comparative Analysis of Classical Test Theory and Item Response Theory Based Item Parameter Estimates of Senior School Certificate Mathematics Examination. *European Scientific Journal, ESJ*, *12*(28), 263. https://doi.org/10.19044/esj.2016.v12n28p263.

Pangalila, T. (2017). Interaksi Sosial Dosen Dan Mahasiswa Dalam Proses Perkuliahan Di Jurusan PPKN FIS UNIMA. *PKn Progresif*, *12*(2), 699–706. https://jurnal.fkip.uns.ac.id/index.php/progresif/article/view/11257.

Pangesti, F., Fauzan, F., & Risnawati, R. (2020). Kualitas butir soal try out uji pengetahuan dalam memprediksi tingkat kelulusan mahasiswa PPG. *Jurnal Pendidikan Profesi Guru*, *1*(2), 91–98. https://doi.org/10.22219/jppg.v1i2.13503.

Prawiki, S. M., & Helendra, H. (2022). Analisis Kualitas Butir Soal Ujian Akhir Semester Ganjil Tahun Pelajaran 2020 / 2021 Mata Pelajaran Biologi Kelas X SMA Negeri 1 Teluk Sebong Analysis of Quality The Question Final Exam Odd Semester 2020 / 2021 Biology Class X SMA Negeri 1 Teluk Sebong. *Biodidaktika: Jurnal Biologi Dan Pembelajarannya*, *17*(2). https://doi.org/10.30870/biodidaktika.v17i2.16493.

Purniasari, L., Masykuri, M., & Ariani, S. R. D. (2021). Analisis Butir Soal Ujian Sekolah Mata Pelajaran Kimia SMA N 1 Kutowinangun Tahun Pelajaran 2019/2022 Menggunakan Model Iteman dan Rasch. *Jurnal Pendidikan Kimia*, *10*(2), 205–214. https://doi.org/10.20961/jpkim.v10i2.48244.

Qurrota, A. A. S., Siskawati, F. S., & Irawati, T. N. (2022). Analisis Kelayakan Butir Soal pada Media INTERMATHLY (Interesting Mathematic Monopoly). *Jurnal Cendekia : Jurnal Pendidikan Matematika*, *6*(1), 634–654. https://doi.org/10.31004/cendekia.v6i1.1181.

Rahman, A., Naldi, W., & Arifin, A. R, F. M. (2021). Analisis UU Sistem Pendidikan Nasional Nomor 20 Tahun 2003 Dan Implikasinya Terhadap Pelaksanaan Pendidikan Di Indonesia. *JOEAI (Journal of Education and Instruction)*, *4*(20), 6. https://journal.ipm2kpe.or.id/index.php/JOEAI/article/view/2010/1419.

Riswanda, H., & Burhan, N. (2022). Analisis butir soal latihan penilaian akhir semester ganjil mata pelajaran bahasa Indonesia kelas VIII SMPN 1 Bambanglipuro Bantul menggunakan program ITEMAN. *KEMBARA: Jurnal Keilmuan Bahasa, Sastra, Dan Pengajarannya*, *8*(1), 160–180. https://doi.org/10.22219/kembara.v8i1.20530.

Rukiyati. (2019). Tujuan Pendidikan Nasional Dalam Perspektif Pancasila. *Humanika*, *19*(1), 56–69. https://journal.uny.ac.id/index.php/humanika/article/download/30160/13136.

Sanaky, M. M. (2021). Analisis Faktor-Faktor Keterlambatan Pada Proyek Pembangunan Gedung Asrama Man 1 Tulehu Maluku Tengah. *Jurnal Simetrik*, *11*(1), 432. https://doi.org/10.31959/js.v11i1.615.

Sanusi, R. N. A., & Aziez, F. (2021). Analisis Butir Soal Tes Objektif dan Subjektif untuk Keterampilan Membaca Pemahaman pada Kelas VII SMP N 3 Kalibagor. *Metafora: Jurnal Pembelajaran Bahasa Dan Sastra*, *8*(1), 99. https://doi.org/10.30595/mtf.v8i1.8501.

Srika Ningsih Pasi; Yusrizal. (2018). Analisis Butir Soal Ujian Bahasa Indonesia Buatan Guru MTSN di Kabupaten Aceh Besar. *Master Bahasa*, *6*(2), 195–202. https://doi.org/10.24173/mb.v6i2.11666.

Sugiyono. (2019). Metode Penelitian Pendidikan. In *Bandung:Alfabeta*.

Suharman. (2018). Tes Sebagai Alat Ukur Prestasi Akademik. *Jurnal Ilmiah Pendidikan Agama Islam*, *10*(1), 93–115. http://ejournal.staindirundeng.ac.id/Index.Php/Tadib/Article/View/138.

Sujana, I. W. C. (2019). Fungsi Dan Tujuan Pendidikan Indonesia. *Jurnal Pendidikan Dasar*, *4*(1), 29. https://doi.org/10.25078/aw.v4i1.927.

Sumiati, A., Widiastuti, U., & Suhud, U. (2018). Workshop Teknik Menganalisis Butir Soal dalam Meningkatkan Kompetensi Guru di SMK Cileungsi Bogor. *Jurnal Pemberdayaan Masyarakat Madani (JPMM)*, *2*(1), 136–153. https://doi.org/10.21009/jpmm.002.1.10.

Supiyansyah, H., Kusumah, I. H., & Berman, E. T. (2017). Analisis Kualitas Soal Ulangan Akhir Semester Genap pada Mata Pelajaran Produktif Program Keahlian Teknik Kendaraan Ringan. *Journal of Mechanical Engineering Education*, *4*(1), 52. https://doi.org/10.17509/jmee.v4i1.7441.

Supriyati, Y., & Dudung, A. (2019). *Penilaian Kelas*. Karima (Karya Ilmu Media Aulia).

Syachlani, A., & Setyorini, D. (2021). Pengembangan Instrumen Hasil Belajar Matematika Siswa (Tes Pilihan Ganda). *Jurnal Akrab Juara*, *6*(3). https://doi.org/10.58487/akrabjuara.v6i3.1523.

Syaifudin. (2020). Validitas dan Reliabilitas Instrumen Penilaian Pada Mata Pelajaran Bahasa Arab. *Cross-Border:Jurnal Kajian Perbatasan Antarnegara*, *3*(2), 106–118. http://www.journal.iaisambas.ac.id/index.php/Cross-Border/article/view/553.

Tarmizi, P., Setiono, P., Amaliyah, Y., & Agrian, A. (2021). Analisis Butir Soal Pilihan Ganda Tema Sehat Itu Penting Kelas V SD Negeri 04 Kota Bengkulu. *ELSE (Elementary School Education Journal) : Jurnal Pendidikan Dan Pembelajaran Sekolah Dasar*, *4*(2), 124. https://doi.org/10.30651/else.v4i2.7090.

Utami, S. P. T., Juidah, I. M., Eko, S. S., & Yuda, R. K. S. (2020). Analisis Butir Soal Ujian Akhir Semester. *Ournal of Elementary Education*, *2*(2), 274–284. https://journal.unnes.ac.id/sju/index.php/jee/article/view/7488.

Vincent, W., & Shanmugam, S. K. S. (2020). The Role of Classical Test Theory to Determine the Quality of Classroom Teaching Test Items. *Pedagogia : Jurnal Pendidikan*, *9*(1), 5–34. https://doi.org/10.21070/pedagogia.v9i1.123.

Warju, W., Ariyanto, S. R., Soeryanto, S., & Trisna, R. A. (2020). Analisis Kualitas Butir Soal Tipe Hots Pada Kompetensi Sistem Rem Di Sekolah Menengah Kejuruan. *Jurnal Pendidikan Teknologi Dan Kejuruan*, *17*(1), 95. https://doi.org/10.23887/jptk-undiksha.v17i1.22914.

Widayanti, W., Bistari, & Suparjan. (2021). Analisis Butir Soal Pilihan Ganda Penilaian Tengah Semester Pada Pembelajaran Tematik Kelas V Sekolah Dasar Negeri 39 Pontianak Kota. *Jurnal DIDIKA : Wahana Ilmiah Pendidikan Dasar*, *7*(2). https://doi.org/10.29408/didika.v7i2.4370.

Widiyanto, D., & Istiqomah, A. (2020). Evaluasi Penilaian Proses Dan Hasil Belajar Mata Pelajaran PPKn. *Citizenship Jurnal Pancasila Dan Kewarganegaraan*, *8*(1), 51–61. https://doi.org/10.25273/citizenship.v8i1.5385.

Widiyanto, J. (2018). Evaluasi Pembelajaran (Sesuai dengan Kurikulum 2013) : Konsep, Prinsip & Prosedur. *Unipma Press*, 257.

Widyawati, R. (2017). Evaluasi pelaksanaan program inklusi sekolah dasar. *Kelola: Jurnal Manajemen Pendidikan*, *4*(1), 109–120. https://doi.org/10.24246/j.jk.2017.v4.i1.p109-120.

Yusup, F. (2018). Uji Validitas Dan Reliabilitas Instrumen Penelitian Kuantutatif. *Jurnal Tarbiyah: Jurnal Ilmiah Kependidikan*, *7*(1), 17–23. https://doi.org/10.21831/jorpres.v13i1.12884.

Zahiroh, U., & Ritonga, P. S. (2021). Analisis Kualitas Butir Soal Pilihan Ganda Mata Pelajaran Kimia Pada Ujian Akhir Semester (Uas) Kelas Xi Man 2 Kepulauan Meranti. *Jedchem (Journal Education and Chemistry)*, *3*(1), 11–20. https://doi.org/10.36378/jedchem.v3i1.780.

Zainal Arifin. (2019). *Evaluasi Pembelajaran, Prinsip, Teknik, dan Prosedur*. PT Remaja Rosdakarya.

Zainal, N. F. (2020). Pengukuran, Assessment dan Evaluasi dalam Pembelajaran Matematika. *Laplace : Jurnal Pendidikan Matematika*, *3*(1), 8–26. https://doi.org/10.31537/laplace.v3i1.310.