Riwayat: Educational Journal of History and Humanities , 6(3) , 2023, pp. 2056-2068
P-ISSN: 2614-3917, E-ISSN: 2775-5037
DOI: https://doi.org/10.24815/jr.v6i3.34299

Item Analysis of Arabic Midterm Exam Class X SMK Muhammadiyah 4 Yogyakarta Based on Classical Theory with R Program

Syaikha Dziyaulhaq Zein[1]✉, Ana Taqwa Wati[2], Anugrah Arya Bakti[3]
[1]UIN Sunan Kalijaga Yogyakarta, Indonesia
[2]Universitas Muhammadiyah Yogyakarta, Indonesia
[3]Universitas Negeri Yogyakarta, Indonesia
Correspondence Author: zezesyaikha@gmail.com✉

Article history
Received : 2023-03-13
Accepted : 2023-07-25
Published : 2023-08-31

Keywords: Item Analysis, Distinguishing Power, Level of Difficulty, RStudio Program

Abstrak: Penelitian ini menghasilkan hasil analisis butir soal Ujian Tengah Semester Bahasa Arab di SMK Muhammadiyah 4 Yogyakarta menyatakan bahwa laporan tersebut data diperoleh dari hasil tes matematika dengan 25 butir soal dan 88 peserta sebagaimana yang ditampilkan pada lampiran. Dari perhitungan parameter tingkat kesulitan pada software RStudio didapatkan hasil yang baik sebagaimana pada tabel 13. Dari 25 butir soal Ujian Tengah Semester Bahasa Arab atau instrumen yang telah dianalisis dengan menggunakan software RStudio, dihasilkan bahwa nilai parameter pada tingkat kesulitan untuk tiap butir. Didasarkan pada tabel tiap butir yang sebanyak 5 butir soal tergolong dalam kategori "sulit", Sebanyak 11 burit soal tergolong dalam kategori soal "sedang", dan 9 butir sisanya termasuk dalam kategori tingkat kesulitan yang "rendah atau mudah". Penelitian ini juga menggunakan jenis penelitian kuantitatif. Dan untuk menyelidiki butir-butir soal yang diujikan kepada para peserta didik kelas 10 Sekolah Menengah Kejuruan (SMK) Muhammadiyah menggunakan jenis pengelompokan data secara kuantitatif dengan analisis butir-butir soal Ujian Tengah Semeter Genap, untuk mengetahui kualitas produk yang digunakan maka diperlukan rentangan data kuantitatif berupa data mentah dan nilai respons dari soal-soal pilihan ganda yang akan diteliti.

Abstract: This study produces the results of the analysis of Arabic Language Midterm Exam items at SMK Muhammadiyah 4 Yogyakarta stating that the report data obtained from the results of mathematics tests with 25 items and 88 participants as shown in the appendix. From the calculation of the difficulty level parameters in the RStudio software, good results are obtained as in table 13. Of the 25 items of the Arabic Language Midterm Exam or instruments that have been analyzed using RStudio software, it is found that the parameter value at the difficulty level for each item. Based on the table for each item, 5 items are classified as "difficult", 11 items are classified as "medium", and the remaining 9 items are categorized as "low or easy" difficulty. This study also uses a type of quantitative research. And to investigate the items tested to 10th grade students of Muhammadiyah Vocational High School (SMK) using a quantitative type of data grouping by analyzing the items of the Even Mid-Semester Exam, to determine the quality of the product used, a range of quantitative data is needed in the form of raw data and response values from multiple choice questions to be studied.

Available online at
http://jurnal.unsyiah.ac.id/riwayat/

INTRODUCTION

Language is a tool that is very useful for the survival of humans in establishing communication. In Indonesia, various languages and types of languages are actively used by the people. Starting from this, the fact is that many people like to pursue and study foreign languages, in order to increase and hone the skills of the learning community so that they can increase

creativity in communicating in the international arena. Foreign languages are also useful so that modern society, especially language learners, get great opportunities to work and build relationships with foreigners. Likewise in the world of lectures, learning foreign languages, especially Arabic, has spread widely and has its own attraction for students.

In a language learning process, language analysis is needed which aims to view and criticize a product brought by an environment that will be tested on students or students. Language analysis that can be used by conducting oral and written tests. A test is a means used to measure an assessment in the form of providing certain exercises or tasks to produce a score that is appropriate or inappropriate. A good test must meet good criteria such as adequate validity and reliability. Is this test suitable for students who will take the Mid-Semester Examination (UTS) or not? The test which is the scope of the assessment provides its own weight in determining a criterion that has been determined together.

In this modern era, in order to see the quality of Arabic language learning that has been carried out by each student at Muhammadiyah 4 Yogyakarta High School or Vocational High School (SMK), it is necessary to analyze the test questions, whether the questions were made for Midterm exam tests are suitable for reuse or not. As explained by Thua'aimah in the article by Halomoan et al, a test is a collection of writing consisting of various questions asked by the (language) teacher so that they can be answered by (the students). Therefore, the test in Arabic itself is useful for honing and understanding Arabic language skills for speakers and other speakers, which consists of several stages. Language tests are closely related to language learning that will be or has been carried out, both local and foreign languages. This test functions whether language learners or students have mastered the language or not. Are students able to implement it or is it just a matter of memory?

According to Arikunto, the activity of analyzing questions is a form of systematic procedure, meaning that an activity of analyzing products is carried out regularly, which will have a big impact in providing information on the test object being studied. In general, analysis activities consist of two forms, namely qualitative analysis and quantitative analysis.

The aim of this research is to be able to analyze the quality of the questions in the Class 10 Even Semester Midterm Examination at Muhammadiyah 4 Yogyakarta Vocational High School (SMK) whether they are in accordance with Arabic language aspects or not. Likewise, the Class 10 Even Semester Midterm Exam questions at SMK Muhammadiyah 4 Yogyakarta are expected to be able to inspire students and exam takers so they can develop their Arabic language skills. Apart from that, this research can provide input as to whether the questions for the Grade 10 Even Semester Midterm Exam at SMK Muhammadiyah Yogyakarta that have been presented are considered worthy of being tested and meet the IRT standard questions or not. Therefore, this problem deserves research so that it becomes additional value for students and the institution. Research similar or relevant to this research was presented by researcher Nurjanah with "Article title Analysis of Multiple Choice Question Items from Linguistic Aspects" which was published in the Journal of Educational Sciences. The article explains the topic in the form of an analysis of the quality of the Class VIII Odd Semester Examination questions for Junior High School (SMP) according to linguistic aspects.

METHOD

The research method used is Quantitative Research Method with a descriptive approach. The analysis used to investigate the question items tested on grade 10 Muhammadiyah Vocational High School (SMK) students uses a quantitative type of data grouping with analysis of the Even Semester Midterm Examination question items. To determine the quality of the product used, a range of quantitative data is needed in the form of raw data and response values from the multiple choice questions to be studied. The answer sheets

obtained by the researchers amounted to 85 items of raw data/processed data which in the next stage the author will explain into grading using the IRT (Item Response Theory) system. Quantitative analysis is carried out as a step to determine and analyze the level of difficulty by using the proportion of correct question answers (Proportion correct) and then determining the range of different power indices.

RESULTS AND DISCUSSION

In the context of learning Arabic, good Arabic learning activities encourage the achievement of learning goals by prioritizing assessment aspects such as cognitive, psychomotor and affective aspects. Because quality Arabic language learning can improve the thinking system and concentration power of language learners in improving their skills. At the beginner learner stage, listening skills are considered quite difficult for those who have not been trained to hear Arabic vocabulary in their daily activities. The Arabic language test can be defined as a learning testing activity in the field of Arabic in the form of questions that have been grouped based on aspects of proficiency as well as levels of difficulty and ease.

Language skills cannot be heard, seen or read because of their abstract nature, even though language skills are always behind the use of language. Language skills on the other hand are concrete, because they refer to the actual system of language use, both spoken and written. All of these things are the purpose of a language test. Test taking refers to the instructions on the test given. Everyone who studies and teaches Arabic should understand and master the knowledge that will be taught well. To determine the quality of the product to be expanded, teachers need a lot of quantitative value information to explain the quality of the product. The material expert's assessment of product quality includes learning aspects and the material presented.

The Arabic Language Mid-Semeter Examination is an activity in the form of an academic test to test the proficiency level of Arabic language learners. The Mid-Semester Exam at Muhammadiyah School 4 Yogyakarta is usually held in the classroom, and students who are carrying out field work practice are given relief to take the test using an *e-learning application* that has been facilitated by the agency. This test is a graduation requirement for students to advance to class level. The language test tested on the Even Midterm Exam questions consists of 25 multiple choice questions. Some questions contain elements of testing vocabulary and Arabic grammar that students have learned.

QUESTION ITEM ANALYSIS

Basically, an analysis activity is an investigation of an event object that is useful for finding out the actual situation. There are several types of question analysis, qualitative question analysis and quantitative question analysis. These two analyzes were carried out to find out whether the questions tested on language learners functioned well in terms of quality and quantity or not. According to Arikunto, item analysis can help language learners and teachers identify the form of questions being tested, get an idea of the style and grammar used, as well as topics that frequently appear in language exams.

Analysis of the questions was carried out to determine the cognitive aspects of all students taking this subject. According to Bloom there are several Bloom's taxonomies including the following; Remembering, Understanding, Applying, Analyzing, Evaluating, Carrying out creative innovations. The steps for writing questions according to Bloom are a) Determining the objectives of the test being held, b) Arranging the Grid, c) Writing the Questions,

d) Reviewing the Questions, e) Trying out the Questions, f) revising and assembling the questions into a test.

Test question analysis is divided into two, namely qualitative question analysis and quantitative question analysis. To arrive at correct and valid question items, it is necessary to carry out theoretical analysis activities with reference to item difficulty, differentiating power, and question functions. Quality analysis in the form of testing questions before being tried out on students in terms of material, construction, language and culture in accordance with established assessment rules. Aspects of the material used in multiple choice questions are: a) Questions according to learning indicators, b) Functional distractors, c) Only one and correct answer key. Meanwhile, the construction aspect is in the form of a) short and concise questions, b) the main formulation of the question is in accordance with the answer, c) the question does not provide direct instructions to the answer key, d) the question must be free from negative elements, e) pictures and graphs must be functional, f) answer choices must avoid the statement "all the answer choices above are correct". Meanwhile, cultural and linguistic aspects that must be emphasized in multiple choice questions are a) the language used is clear and in accordance with the rules, b) the language is communicative, c) reducing regional languages, d) answer choices are prohibited from using repetitions of the same words.

MIDDLE SEMESTER EXAMINATION TEST

According to Purwanti (2014) in Lailatul Qomariah's article, a test (إختبار)is an instrument to measure students' proficiency level. The test usually contains several questions that have been designed by the teacher creatively and meet standard questions so that they can test the students' abilities that will be achieved and aim to represent the achievement of learning that has been completed during the specified period of time. Arifin (2012), Arikunto (2013) and Munthe (2015) and Ma'arif expressed their opinions regarding tests which are tools or devices used to test learning activities that have been implemented.

Overview of the RStudio Program Application

Analysis of test instruments in evaluation in the education sector can be carried out using two approaches. The first approach, namely the approach that is most widely and commonly applied in the field of education to date, is classical test theory (CTT). The aspects that really determine the quality of the items in the classical test theory approach are the level of difficulty and distinguishing power of the questions. However, the characteristics of the items produced by classical test theory are inconsistent (change) depending on the ability of the test taker.

R PROGRAM

The R program is a series of integrated software facilities for manipulating, calculating, and displaying graphics of data. When using this program, it has many advantages. Some of them are effective data handling and deviation facilities, graphical facilities for data analysis and display, and the programming language is well developed and simple. From there it can be concluded that R is a good tool for newly developing interactive data analysis methods. The development of the R program itself is very fast and expands with a large collection of packages. However, most programs written in R are essentially ad hoc, written for a single data analysis. R can be considered an implementation of the S language developed at Bell Laboratories by Rick Becker, John Chambers and Allan Wilks, and also forms the basis of the S-Plus system (Shah, 2013).

Since its first release in 1992 by Ross Ishaka and Rob Gentleman, two statisticians from the University of Auckland in New Zealand, the R language has grown rapidly in terms of functionality and use. This can be seen by looking at the resources provided online by the Comprehensive R Archive Network (CRAN) (https://cran.r-project.org/). This website provides R project documentation, precompiled binary distributions of the R language for various operating systems, source code, R manuals, and even a scientific journal called The R Journal (Krotov, 2017). In analyzing test items, the R program offers a package that can be used to analyze test items based on classical test theory, namely "CTT" and "psychometric". The following is an initial display of the RStudio software.

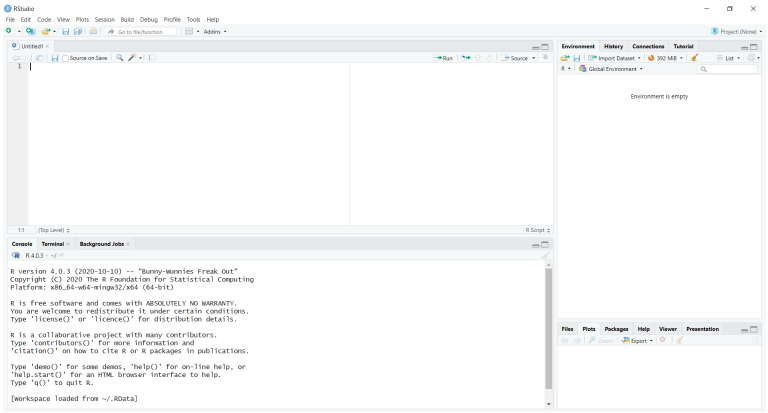


Figure 1. Rstudio software display

Analysis of Question Items based on Classical Test Theory

Item analysis can be carried out using classical and modern approaches. The classical approach is usually called Classical Test Theory (CTT). In the classical test theory approach, the true score (true score = T) and error score (error score = E) are theoretical constructs that cannot be observed. Only observed score (observed score = X) can be obtained, and observed score = true score + error (X = T + E). If we talk about actual scores, it is important to remember that the actual score, namely the average score obtained from independent repetition of the test using the same test, is purely theoretical. This score does not completely show the true characteristics of the test taker unless the test has perfect validity, namely that the test measures exactly what the content being measured is.

Quantitative analysis according to the classical test theory approach produces item characteristics which include the level of difficulty (p), distinguishing power (d), and distractor effectiveness. Apart from that, with quantitative analysis using a classical theoretical approach, it is also possible to determine the reliability of test questions and the standard error of measurement. To see the level of difficulty, differentiating power and effectiveness of distractors, an analysis of each test item is carried out, while the reliability and standard measurement error can be seen by analyzing the test questions as a whole. In classical item analysis, the level of difficulty is one of the parameters of the item, which is symbolized (Pi), namely the ratio between the correct answer and the number of answerers to the item. The formulation of the difficulty level of the questions is as follows.

$$P_i = \frac{n}{N}$$

Where P_i is the difficulty level of item i, n is the number of students who answered the question item correctly, and N is the number of students who answered the question item. The level of difficulty ranges between zero and one. An item is sometimes categorized into extreme difficulty, namely if the p value is close to zero and extreme easy if the p value is close to one. According to Fernandes (1984), questions that produce an average score of around 50% of the maximum score can be said to have the right level of difficulty. Meanwhile, Thomas and Dawson (1972) explained that items with a difficulty level of 0.25 - 0.75 were said to be good.

The distinguishing power or differentiability of a test item functions to determine whether or not a test item can differentiate between groups in the aspects measured according to the differences that exist in that group. The purpose of examining discriminating power is to see the ability of certain test items to differentiate between high-ability test takers and low-ability test takers. Differential power calculations can be done using the correlation method, namely point biserial correlation and biserial correlation. Point biserial correlation and biserial correlation are product moment correlations applied to data, the variables being correlated are each different from each other. According to Crocker & Algina (1986) the point biserial coefficient is determined by the following formula.

$$\rho_{pbis} = \frac{\mu_+ - \mu_\tau}{\sigma_\tau}\sqrt{\frac{p}{q}}$$

This ρ_{pbis} is the point biserial correlation, μ_+ which is the average of participants who answered the question item correctly, μ_τ which is the average of the total score, σ_τ which is the standard deviation of the total score, p is the proportion of participants who answered correctly, and q is 1-p. Meanwhile, the discriminating power index with biserial correlation can be calculated using the following formula.

$$\rho_{bis} = \frac{\mu_+ - \mu_\tau}{\sigma_\tau}\left(\frac{p}{Y}\right)$$

As in the previous formula, with the addition ρ_{bis} of biserial correlation and Y is the ordinate p in the normal distribution. The differential power index of test items can be used as a consideration for whether an item is good or not. Good test items are items that have a different power index of more than 0.2. as stated by Fernandes (1984). Meanwhile, Ebel (1972) explains that an item is said to be of quality if the discrimination index or differentiating power is at least 0.41. Ebel (1965) offered the following guidelines for item interpretation based on his item discrimination index, which are as follows:

Table 1. Discrimination Index Categories

Indeks Diskriminasi	Interpretasi
0,40 atau lebih	Item berfungsi sangat baik
0,30 – 0,39	Item sudah baik, mungkin diperlukan revisi kecil
0,20 – 0,29	Item masih meragukan dan perlu direvisi
Kurang dari 0,20	Item harus dibuang atau direvisi total

An important thing that must also be considered in empirically analyzing the questions is the ability of the distractors or alternative answers provided to attract test takers to choose them. Don't let none of the test takers choose the alternative answers provided. Fernandes (1984), citing Brawn's opinion, explains that a distractor is said to be good if it is chosen by at least 2% of all participants. Meanwhile, Nitko (1996) said that a distractor is said to be functional if at least one test taker from the low group chooses it. There must be more voters from the lower group than the upper group. Distractors can also be said to function when test takers (students) from the upper group can differentiate between distractors and answer keys so that more people choose answer keys than choose distractors.

Thus it is clear that to assess the quality of test items it is not enough to just pay attention to the level of difficulty and differentiating power of the test items in question. Assessment of the quality of test items must also look at the function of the answer choices, especially the distractors, that is, they must appear to be the correct answer for subjects from low ability groups. Otherwise it should appear to be an incorrect answer to subjects from the high ability group. Even if a test item is too difficult or too easy, if (1) the differentiating power of the test item, and (2) the distribution of answers meet the criteria, then the test item can still be accepted as a good test item. The criteria in question are the discriminating power index for RBIS test items > 0.3, and the discriminating power index for negative answer choices except the key (Kartowagiran, 2009). In general, the criteria for selecting multiple choice questions or dichotomous data are as follows.

Table 2. Question Selection Criteria in Classical Test Theory

Kriteria	Koefisien	Keputusan
Tingkat Kesukaran	0,30 s.d. 0,70 (sedang)	Diterima
	0,10 s.d. 0,29 atau 0,70 s.d. 0,90 (sukar atau mudah)	Direvisi
	< 0,10 atau > 0,90 (sangat sukar atau sangat mudah)	Ditolak
Daya Pembeda	> 0,3	Diterima
	0,10 s.d 0,29	Direvisi
	< 0,10	Ditolak
Proporsi Jawaban	> 0,05	Berfungsi baik

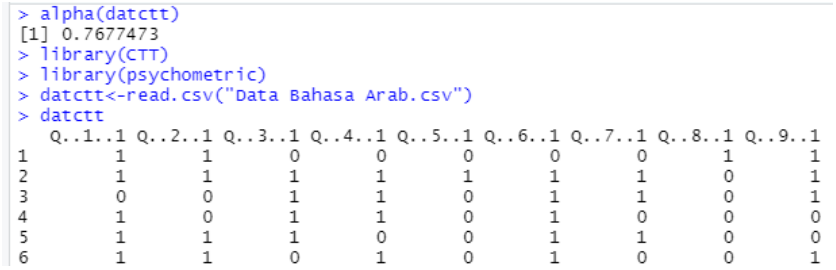
Results of Analysis of Mid-Semester Exam Questions

The data used in this report is dichotomous data on mathematics test results with 25 questions and 88 participants as in the attachment.

RStudio

Analysis using RStudio begins by installing the "CTT" and "psychometric" packages. Then proceed with data input and write the syntax needed for item analysis based on classical test theory. The syntax for this report is written in Appendix 5. After all the desired syntax has been written, run the program by clicking "Run". The following is

an example of the statistical results obtained for question number 1.



```
> alpha(datctt)
[1] 0.7677473
> library(CTT)
> library(psychometric)
> datctt<-read.csv("Data Bahasa Arab.csv")
> datctt
  Q..1..1 Q..2..1 Q..3..1 Q..4..1 Q..5..1 Q..6..1 Q..7..1 Q..8..1 Q..9..1
1       1       1       0       0       0       0       0       1       1
2       1       1       1       1       1       1       1       0       1
3       0       0       1       1       0       1       1       0       1
4       1       0       1       1       0       1       0       0       0
5       1       1       1       0       0       1       1       0       0
6       1       1       0       1       0       1       0       0       1
```

Figure 2. Analysis output in RStudio

From one of the RStudio outputs, a reliability value of 0.767 was obtained, which means that it can be concluded that the reliability of an instrument with 88 participants and 25 items is in the high category.

Different Power

The following table shows the results of the analysis of Different Power Parameters

Table 12 . Different Power Parameters with RStudio

Item No	Different Power		
	Coefficient	Information	Decision
1	0.357	Good	Accepted
2	0.392	Good	Accepted
3	0.142	Pretty good	Revised
4	0.357	Good	Accepted
5	0.178	Pretty good	Revised
6	0.321	Good	Accepted
7	0.321	Good	Accepted
8	-0.142	Not good	Rejected
9	0.321	Good	Accepted
10	0.607	Good	Accepted
11	0.464	Good	Accepted
12	0.571	Good	Accepted
13	0.500	Good	Accepted
14	0.464	Good	Accepted
15	0.285	Pretty good	Revised
16	0.464	Good	Accepted
17	0.250	Pretty good	Revised
18	0.285	Pretty good	Revised
19	0.428	Good	Accepted

20	0.464	Good	Accepted
21	0.107	Pretty good	Revised
22	0.357	Good	Accepted
23	0.178	Pretty good	Revised
24	0.285	Pretty good	Revised
25	0.607	Good	Accepted

Calculation of the differential power for each item in this instrument which was carried out using RStudio software produces a summary of the differential power values as in the following table. From the 25 questions on the instrument that were analyzed using RStudio software, different power parameter values were produced for each item. Based on the table, it is obtained as much as

1. A total of 16 items are included in the **"good" category,**
2. A total of 8 items are included in the **"fairly good" category,** and
3. 1 item in the **"not good" category.**

Level of Difficulty n

The following table shows the results of the Difficulty Level Parameter analysis

Table 13. Difficulty Level Parameters with RStudio

Item No	Difficulty Level		
	Coefficient	Information	Decision
1	0.666	Currently	Accepted
2	0.559	Currently	Accepted
3	0.321	Currently	Accepted
4	0.761	Easy	Revised
5	0.250	Difficult	Revised
6	0.857	Easy	Revised
7	0.238	Difficult	Revised
8	0.226	Difficult	Revised
9	0.357	Currently	Accepted
10	0.619	Currently	Accepted
11	0.559	Currently	Accepted
12	0.571	Currently	Accepted
13	0.797	Easy	Revised
14	0.750	Easy	Revised
15	0.702	Easy	Revised
16	0.773	Easy	Revised
17	0.904	Easy	Revised
18	0.690	Currently	Accepted
19	0.690	Currently	Accepted
20	0.357	Currently	Accepted

21	0.214	Difficult	Revised
22	0.702	Easy	Revised
23	0.190	Difficult	Revised
24	0.738	Easy	Revised
25	0.642	Currently	Accepted

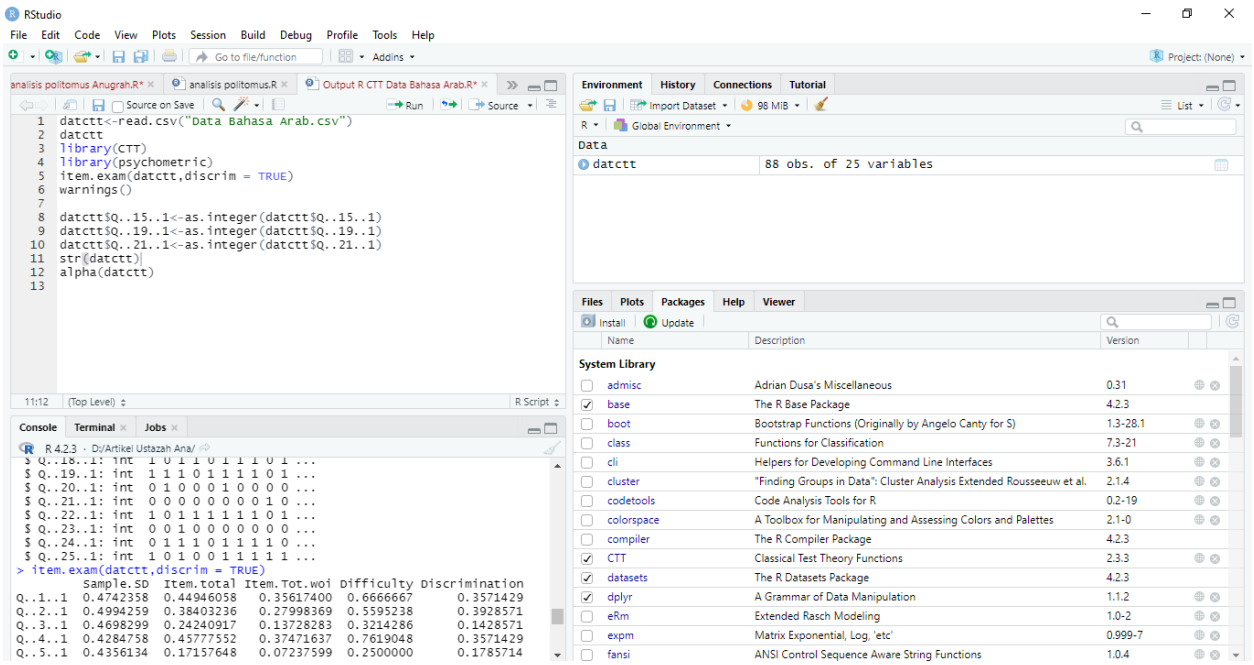
Calculation of the difficulty level parameters in the RStudio software resulted in the results as in Table 13. From the 25 items on the instrument that were analyzed using the RStudio software, a difficulty level parameter value was produced for each item. Based on the Item Table

1. A total of 5 items are included in the **"Difficult" category** .
2. A total of 11 items are included in the **"Medium" question category.**
3. and the remaining 9 items fall into the "low or easy" difficulty level category **.**

The categorization of item goodness based on difficulty level parameters refers to the question selection criteria based on classical test theory in Table.

Attachment

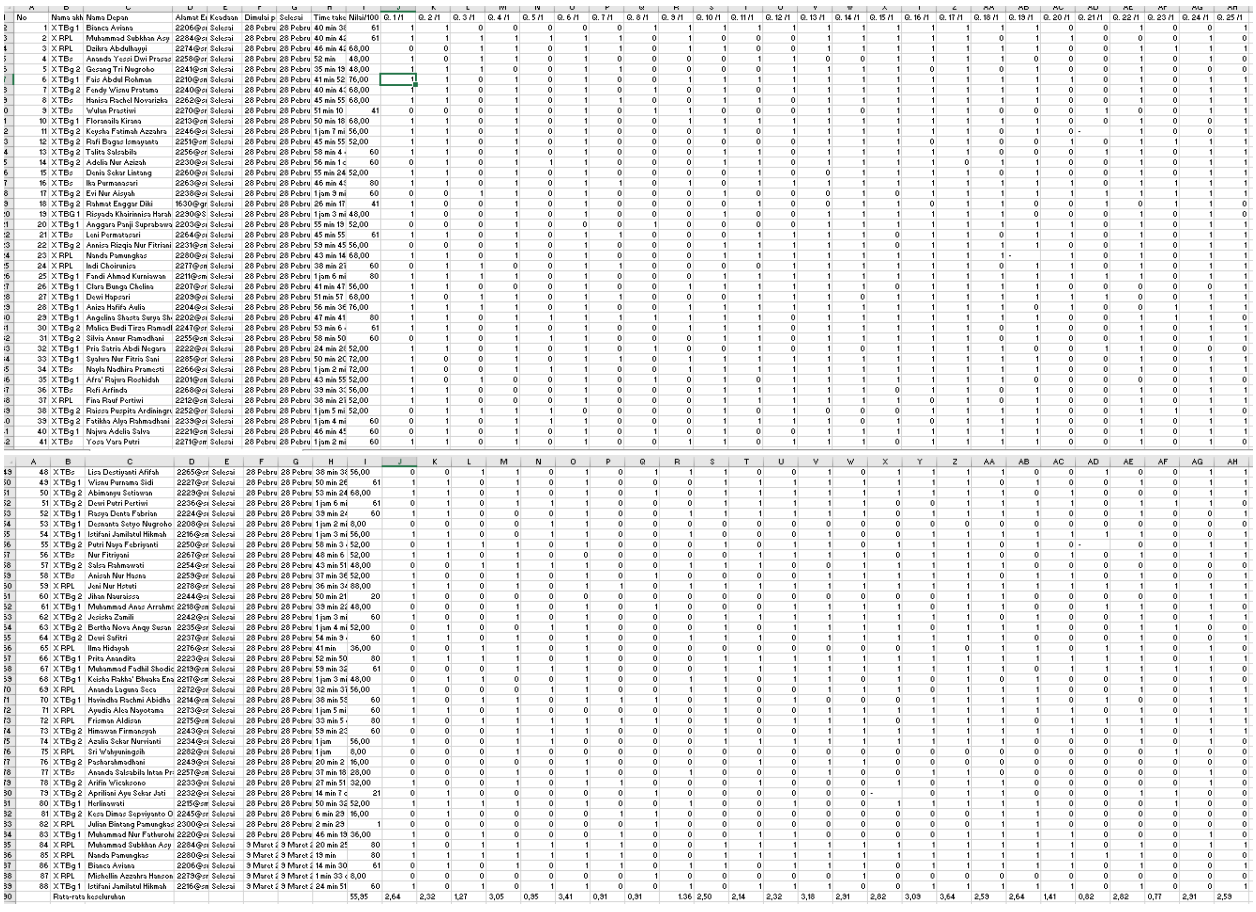
1. R Program Analysis



2. Analysis Results of a Program

```
> item.exam(datctt,discrim = TRUE)
         Sample.SD  Item.total Item.Tot.woi Difficulty Discrimination
Q..1..1  0.4742358  0.44946058   0.35617400  0.6666667       0.3571429
Q..2..1  0.4994259  0.38403236   0.27998369  0.5595238       0.3928571
Q..3..1  0.4698299  0.24240917   0.13728283  0.3214286       0.1428571
Q..4..1  0.4284758  0.45777552   0.37471637  0.7619048       0.3571429
Q..5..1  0.4356134  0.17157648   0.07237599  0.2500000       0.1785714
Q..6..1  0.3520288  0.46170100   0.39449767  0.8571429       0.3214286
Q..7..1  0.4284758  0.26580514   0.17104550  0.2380952       0.3214286
Q..8..1  0.4208766 -0.09677866  -0.19081959  0.2261905      -0.1428571
Q..9..1  0.4820353  0.28549441   0.17937359  0.3571429       0.3214286
Q..10..1 0.4885376  0.54153350   0.45478086  0.6190476       0.6071429
Q..11..1 0.4994259  0.40066056   0.29788765  0.5595238       0.4642857
Q..12..1 0.4978439  0.49566538   0.40197452  0.5714286       0.5714286
Q..13..1 0.4041878  0.61247496   0.54928432  0.7976190       0.5000000
Q..14..1 0.4356134  0.51472944   0.43538658  0.7500000       0.4642857
Q..15..1 0.4599571  0.47888921   0.39124072  0.7023810       0.2857143
Q..16..1 0.4208766  0.62295391   0.55818173  0.7738095       0.4642857
Q..17..1 0.2953066  0.66688880   0.62655141  0.9047619       0.2500000
Q..18..1 0.4650739  0.28995432   0.18791005  0.6904762       0.2857143
Q..19..1 0.4650739  0.37923644   0.28238991  0.6904762       0.4285714
Q..20..1 0.4820353  0.45203281   0.35733966  0.3571429       0.4642857
Q..21..1 0.4127903  0.19064294   0.09712891  0.2142857       0.1071429
Q..22..1 0.4599571  0.49694428   0.41103088  0.7023810       0.3571429
Q..23..1 0.3950352  0.15916902   0.06912359  0.1904762       0.1785714
Q..24..1 0.4423118  0.38891791   0.29770378  0.7380952       0.2857143
Q..25..1 0.4820353  0.52996950   0.44315734  0.6428571       0.6071429
```

3. Respondent Data



CONCLUSION

The results of the analysis of Arabic language mid-term exam questions at SMK Muhammadiyah 4 Yogyakarta stated that the report data was obtained from the results of a mathematics test with 25 questions and 88 participants as shown in the attachment. From the calculation of the difficulty level parameters in the RStudio software, good results were obtained as in table 13. From the 25 Arabic Mid-Semester Exam questions or instruments that had been analyzed using the RStudio software, the resulting parameter values were at the level of difficulty for each item. Based on the table for each item, 5 questions are in the "difficult" category, 11 questions are in the "medium" category, and the remaining 9 questions are in the "low or easy" difficulty level category.

REFERENCES

Arabiyah Linnasyiin Jilid, "KOMPETENSI PRODUKTIF BERBAHASA ARAB DALAM BUKU AL- م صلختس ثحبلا Pendahuluan" xx, no. x (1907).

Habibi, Burhan Yusuf and Wakhidati Nurrohmah Putri, "Pengembangan Buku Ajar Bahasa Arab Berbasis Pada Nilai-Nilai Islam-Indonesia Di IAIN Salatiga," Tsaqofiya : Jurnal Pendidikan Bahasa Dan Sastra Arab 3, no. 1 (2021): 26–45, https://doi.org/10.21154/tsaqofiya.v3i1.66.

Crocker, L. M., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.

Dawson, J.B. & Thomas, G.H. (1972). Item analysis and examination statics. Birmingham: The Union of Educational Institutions.

Ebel, R. L. (1965). Measuring educational achievement. Englewood Cliffs, N.J: PrenticeHall

Ebel, R.L. (1972). Essentials of educational measurement. (3rd. ed.) Englewood Cliffts,NJ: Prentice Hall Inc.

Farida. Anna Musyarofah, "Validitas Dan Reliabilitas Dalam Analisis Butir Soal," Al-Mu'Arrib: Journal of Arabic Education 1, no. 1 (2021): 34–44, https://doi.org/10.32923/al-muarrib.v1i1.2100.

Fernandes, H.J. X. (1984). Evaluation of educational program. Jakarta: National Education Planning , Evaluating and Curriculum Development.

Halomoan, H n dkk, 2022. "Taḥlīl Ikhtibār Kafā`ah Al-Lugah Al-'Arabiyyah Li an-Nāṭiqīna Bigairihā Fī Jāmi'Ah Sulṭān Syarīf Qāsim Al-Islāmiyyah Al-Ḥukūmiyyah Riau," LISANIA: Journal of Arabic Education and Literature 6, no. 1: 74–87, https://doi.org/10.18326/lisania.v6i1.74-87.

Ida and Musyarofah. 2018. Karimatus Saidah, "Analisis Bentuk-Bentuk Penilaian Sikap Siswa Sekolah Dasar Di Kota Kediri," Profesi Pendidikan Dasar 1, no. 1 (2018): 80, https://doi.org/10.23917/ppd.v1i1.4244.

Ida and Musyarofah, "Validitas Dan Reliabilitas Dalam Analisis Butir Soal.".

Kartowagiran, B. (2009). Pengantar teori tes klasik (ttk)*). Pengantar Teori Tes Klasik, April, 1–19.

Krotov, V. (2017). A Quick Introduction to R and RStudio® Tutorial. November. https://doi.org/10.13140/RG.2.2.10401.92009

Muhtarom, Yusuf. Article Info, and Article History, "Inovasi Penilaian Pembelajaran Bahasa Arab Berbasis Games KAHOOT Untuk Meningkatkan Keterampilan Menyimak Di Ma`Had IIT Rabbani Bengkulu," n.d.

Nitko, A.J. (1996). Penilaian berkelanjutan berdasarkan kurikulum (PB2K): Kerangka, konsep, prosedur, dan kebijakan (terj. AM. Ahmad) Jakarta: Pusat Pengembangan Agribisnis.

Penulisan Butir Soal, "UNIVERSITAS NEGERI YOGYAKARTA," 2012, 1–33.

Qomariyah, Lailatul. 2022. "Analisis Tingkat Kesukaran Dan Daya Pembeda Butir Soal TOAFL Universitas Hasyim Asy'ari Tebuireng Jombang," Lisanan Arabiya: Jurnal Pendidikan Bahasa Arab 6, no. 1, : 1–18, https://doi.org/10.32699/liar.v6i1.2549.

Qomariyah, "Analisis Tingkat Kesukaran Dan Daya Pembeda Butir Soal TOAFL Universitas Hasyim Asy'ari Tebuireng Jombang."

Shah, N. (2013). An Introduction to R. Practical Graph Mining with R, 0, 27–52. https://doi.org/10.1201/b15352-7

Ramadhan, Rachmad and Fasich Nur Firdaus, 2022. "Analisis Butir Soal Ujian Tengah Semester Bahasa Arab Kelas XII Di SMA Al-Izzah IIBS Malang," Tsaqofiya : Jurnal Pendidikan Bahasa Dan Sastra Arab 4, no. 1,: 126–35, https://doi.org/10.21154/tsaqofiya.v4i1.49.