JEEF (JOURNAL OF ENGLISH EDUCATION FORUM)
e-ISSN 2807-260X
p-ISSN 2807-2480

ITEM ANALYSIS ON ENGLISH SUMMATIVE TEST FOR THE 11th GRADE STUDENTS IN MAN 2 MATARAM THE ACADEMIC YEAR 2022/2023

Khanza Phyllana Azizah[1]*, Sahuddin[2], Ahmad Zamzam,[3]
[1 2 3] English Education Department, Faculty of Teacher Training and Education,
University of Mataram, Indonesia
*Corresponding Author: kphylln@gmail.com

Abstract: On November 30th, 2022, MAN 2 Mataram administered a summative English test to students in the eleventh grade. The test results revealed that the majority of the students performed badly on the test, for unclear reasons. The English teacher, who is also the test maker, should take item analysis into consideration while evaluating the performance of the test items. However, because the teacher is unprepared with item analysis, this study intends to provide comprehensive data on the 3 crucial aspects of item analysis; item dfficulty level, item discriminating power, and effectiveness of item distractors of the test. The 50 MCQs of the English summative test as well as 186 students' worksheets were taken as the samples. Arikunto's (2018) theory was utilized to analyze and describe the data. Through the analysis of item difficulty level, it was revealed that out of 50 items, 3 items (6%) were difficult, 28 items (56%) were medium, and 19 items (38%) were easy. In regard to item discriminating power, 1 item (2%) was very poor, 14 items (28%) were poor, 18 items (36%) were satisfactory, 16 items (32%) were good, and 1 item (2%) was excellent. Finally, out of 200 distractors, the analysis suggested that 108 distractors (54%) were functional and 92 distractors (46%) were nonfunctional. In conclusion, majority of the test consisted of items with a medium level of difficulty (56%), satisfactory discriminating power (36%), and functional distractors (54%).

Keywords: summative test, item analysis, item difficulty level, item discriminating power, effectiveness of distractors

INTRODUCTION

In the language classroom, teachers evaluate students' achievements through assessment. There is a variety of tools of assessment that the teacher can utilize, with administering tests being one of them (Hughes, 2003). In educational practice, tests are methods used to ascertain students' capacities to carry out specific tasks, show mastery of a skill, or demonstrate content knowledge in relation to a given standard, which is typically deemed acceptable or not (Adom et al., 2020). Test results can serve as proof of what was learned and taught, they can also serve as feedback on how well the teaching program is working as a whole, information used to guide decisions about the kind of educational resources and activities that should be made available to students, a diagnosis of strength and weaknesses to determine whether a class as a whole or a specific student is prepared enough to move on to next unit of teaching, a guide to assign students' grades based on their achievement, as well as a method of defining the instructional objectives, instructional resources or materials, and activities depending on the students' needs for language acquisition.

According to Cizek (2010), the summative test may be characterized by two criteria: (1) it is carried out at the conclusion of certain units, and (2) its primary objective is to evaluate the performances of the students or systems. According to Woods (2015), the main distinction between a formative and summative test or assessment is that the former is used to support the educational programs, while the latter is used to assess the overall worth of the instructional programs. As mentioned by Shepard (2019), a summative test must be successful in achieving its primary goal of describing what students know and can perform while simultaneously effectively achieving a secondary goal of providing learning assistance. One of the common tools for conducting a summative test is MCQ (Multiple-choice question). A multiple-choice test is a type of objective test that functions to measure various competencies and is made up of a stem and a few possible answers (alternatives/options) consisting of only one correct answer and the others are called distractors. While it takes the examiner more time and effort to create high-quality MCQs than descriptive questions such as an essay, a big chunk of the curriculum is examined in a short amount of time with less work required from the student. A Multiple-choice question is a useful technique for pinpointing students' strengths and shortcomings as well as giving teachers instructions on how to conduct their lessons.

The goals of an evaluation are frequently not met by MCQs that are poorly prepared. A technique for assessing the efficacy of MCQs is item analysis. According to Arikunto (2018), experienced teachers often still struggle to acknowledge that the test is still not ideal since the test results frequently don't match up with expectations. For example, when virtually all pupils receive failing scores, this indicates that the test items are too difficult, and vice versa. Consequently, it can be concluded that there is something that needs to be studied in the performance of the test. Arikunto (2018) continued by stating that, item analysis, a systematic process that will offer information relating to good, terrible, or awful questions, is one method of interpreting test results. Teachers can use item analysis to gather data on the performance of the test so that future tests can be improved. According to Arikunto (2018), there are three crucial aspects of item analysis—item difficulty (P), item discriminating power (D), and distractor efficacy (DE) or effectiveness of distractors—that may be used to assist teachers in deciding whether or not an item is good. The percentage of test takers who gave the correct response reveals how challenging the item was. The difficulty index rises the easier the item is thought to be. Furthermore, the discrimination index reveals whether the questions were able to differentiate between students with high and low scores. If a student gives an answer that is neither correct nor incorrect, the discriminating index (DI) is 0 (Quaigrain & Arhin, 2017). Another important technique is the analysis of the distractors, which provides information on the individual distractors and the answer to the test item. This technique allows the examiner to modify or remove certain questions from future examinations.

In conclusion, given the importance of evaluating a test, therefore, teachers must ensure that they are creating a high-quality test, which is rather challenging. As Brown (2003) mentioned it takes a lot of work to make a well-constructed test that accurately assesses the proficiency of the students within a particular subject. Ebel and Frisbie (1991) explained that when creating an effective test, a test maker or a teacher should consider the use of items with moderate difficulty levels which can discriminate between high and low performers. One of the methods of evaluating the quality of a test is by conducting an item analysis. Item analysis is a set of procedures for evaluating the quality of test items. When conducting item analysis, the teacher gets information regarding the level of difficulty of an item, the discriminating power, and the effectiveness of distractors, which can help to determine which items can be accepted, revised, and rejected.

RESEARCH METHODS

This research was considered descriptive research which is "concerned with describing the characteristics of a particular individual, or of a group" (Kothari, 2004, p. 37). A descriptive qualitative describes the numerical value that is obtained from quantitative analysis into some criteria or qualities such as good and bad (Hikmawati,2020). The populations were the 50 items of the English summative test and the 400 eleventh-grade students' worksheets in MAN 2 Mataram in the academic year 2022/2023. Simple random sampling was employed to get a sample of 186 students' worksheets. To collect the data, the researcher used documents such as the 50 English test items, key answers, and 186 students' worksheets. The data were then analyzed through Arikunto's (2018) theory and formula. Specifically, the techniques are as follows:

Item difficulty level:

$$P = \frac{B}{JS}$$

P = Item difficulty level
B = number of students with correct answers
JS= total number of students

P	Interpretations
0 – .30	Difficult
.31 – .70	Medium
.71 – 1	Easy

Item discriminating power:

D= PA-PB

D = item discriminating power
PA = Proportion of students in the upper group answering correctly
PB = Proportion of students in the lower group answering correctly

D	Interpretations
Negative	Very poor
0 – .20	Poor
.21 – .40	Satisfactory
.41 – .70	Good
.71 – 1	Excellent

Effectiveness of item distractor:

This is done through comparing the number of students in the upper and lower groups who chose each distractor

Interpretations

NF-D: chosen by <5% of students & chosen more by students in the upper group

F-D: chosen by ≥5% of students & chosen by more students in the lower group

FINDINGS AND DISCUSSION

Findings

In accordance with the research questions, this sub-chapter analyzed and described the quality of each item in terms of the difficulty level, discriminating power, and also effectiveness of distractors through item analysis procedure proposed by Arikunto (2018). Before the analysis, the researcher took 186 worksheets of the 11th-grade students in MAN 2

Mataram in the academic year 2022/2023 as the sample then divided these worksheets into two groups by taking the highest 27% (50 students' worksheets) and the lowest 27% (50 students' worksheets) of the worksheets based on the rank of the scores and referred to these groups as the upper and lower group.

a. Item difficulty level

Table 1. Results of item difficulty level

Item difficulty level (P)	Total	Percentage
Difficult (0 – .30)	3	6%
Medium (.31 – .70)	28	56%
Easy (.71 – 1)	19	38%

The researcher obtained the difficulty level (P) of each item by summing up the number of correct answers of the upper and lower groups and dividing it by the number of students from both groups. The result of difficulty level is also referred to as the difficulty index which starts from 0 to 1.00. To specify the interpretation, the researcher relied on Arikunto's (2018) table of difficulty index interpretation which shows that P with 0 to .30 means the item is difficult, P with .31 to .70 is medium, and P with .71 to 1 is easy. An appropriate difficulty level is between .31 to .70 which is not too easy nor too difficult for the students of the upper and lower groups.

The difficulty level of the English summative test given to the 11th-grade students in MAN 2 Mataram in the academic year 2022/2023 was revealed through item analysis and was divided into three categories: easy (19 items), medium (28 items), and difficult (3 items).
To explain the results, the researcher only mentioned 1 item as the representative of each category, such as:

Difficult:
Item 8: the correct answer for item 8 was option E. The item analysis revealed that there were 9 upper group students who answered correctly whilst only 3 students from the lower group did. The analysis of item difficulty level resulted in .12, which according to Arikunto (2018) belonged to the category of a difficult item.

Medium:
Item 22: the correct answer for item 22 was option C. There were 35 students in the upper group and 20 of those in the lower group answering correctly on this item. The difficulty level resulted in .55, which in accordance with Arikunto (2018), belonged to the category of a medium item.

Easy:
Item 12: the correct answer for this item was option D. The total number of students in upper group answering correctly was 49, meaning that almost everyone did. As for the students in the lower group, there were 38 of them did correctly. The difficulty level of this item was .87, therefore, it was considered easy.

b. Item discriminating power

Table 2. Results of item discriminating power

Discriminating power	Total	Percentage

(D)		
Very poor (negative)	1	2%
Poor (0 – .20)	14	28%
Satisfactory (.21 – .40)	18	36%
Good (.41 – .70)	16	32%
Excellent (.71 – 1)	1	2%

Item discriminating power (D) was obtained by substracting the proportion of the upper group who answered correctly from that of the lower group. The high-performing students should be able to answer items correctly more than the low-performing students do, otherwise the items fail to discriminate the students. the results of item discriminating power were then interpreted using Arikunto's (2018) interpretation that says D with negative result is very poor, 0.00 to 0.20 is poor, 0.21 to 0.40 is satisfactory, 0.41 to 0.70 is good, and 0.71 to 1.00 is excellent.

Through item analysis, the researcher was able to find out the item discriminating power of the English summative test given to 11th-grade students in MAN 2 Mataram in the academic year 2022–2023 which fell into 5 categories: very poor/negative (1 item), poor (14 items), satisfactory (18 items), good (16 items), and excellent (1 item).

To explain the results, the researcher only mentioned 1 item as the representative of each category, such as:

Very poor:

Item 7: this item had a discriminating power -.04, which according to Arikunto (2018) very poorly or failed to discriminate between the upper and lower group students. The content covered in item 7 had to do with deciding the best title for the reading passage. There was an equal number of students in the upper group selecting option A (23 students) and C (23 students), this may indicate that they found both options to be correct or similar. However, the reason this item had a negative index for item discriminating power was because there were more students in the lower group selecting C (the correct answer). We may assume those students answered this item merely by guessing or asking their peers. It is quite common for the students to answer an item correctly even without knowledge because an MCQ provides a wider chance for the students to guess or exchange answers with their friends.

Poor:

Item 8: the discriminating power (D) was 0.12, which according to Arikunto (2018), belonged to the category of a poor item meaning that this item poorly discriminated between the upper and lower groups. This item required the students to determine the antonym of a word. Very small number of students in both groups could barely answer correctly meaning that it was somehow too difficult to answer. By looking at the stem and alternatives, the students might have confused the instruction of the stem with synonyms or another possible factor may be due to the fact they still had a lack of vocabularies. Therefore, this item cannot help the teacher as the test maker to differentiate the ability of students in the upper and lower groups.

Satisfactory:

Item 50: the discriminating power (D) was 0.28, which according to Arikunto (2018), belonged to the category of a satisfactory item meaning that it acceptably discriminated between the upper and lower groups.

Good:
Item 47: the discriminating power (D) was 0.7, which according to Arikunto (2018), belonged to the category of a good item meaning that it discriminated between the upper and lower groups well.
Excellent:
Item 46: the discriminating power (D) was 0.72, which according to Arikunto (2018), belonged to the category of an excellent item meaning that it perfectly discriminated between the upper and lower groups.

c. Effectiveness of item distractors

Table 3. Results of the effectiveness of item distractors

Effectiveness	Total	Percentage
Functional distractors (FD)	108	54%
Nonfunctional distractors (NF-D)	92	56%

The English summative test assessed to the 11th-grade students in MAN 2 Mataram in November 2022 was in the form of a multiple-choice test consisting of 50 items with one correct answer and four distractors per each item, therefore, there were 200 distractors in total. Analyzing the effectiveness of distractors will result in whether a distractor is considered functional or nonfunctional. A functional distractor (F-D) is the one that is selected by at least 5% of the total students from the upper and lower groups and the lower group should be drawn to select it more than the upper group. On the other hand, if a distractor is selected by less than 5% of students or the upper group dominates the selection, then it is a nonfunctional distractor (NF-D). Through item analysis, the researcher found that out of the 200 distractors, there were 108 functional distractors (F-D) and 92 nonfunctional distractors (NF-D).
To explain the results, the researcher only mentioned 3 items as the representative of each category, such as:

a. Item 2: out of the four distractors, only distractor C was functional distractor (F-D) whilst the rest was nonfunctional-distractor (NF-D). Distractors A and D were not selected at all, whereas distractor E was only selected by a single student from the lower group meaning that not even 5% of the group selected it.
b. Item 28: distractor E was the only functional distractor (F-D) as it attracted more students from the lower group. However, distractors A and B were nonfunctional as they received selections from less than 5% of students. Moreover, distractor C attracted the upper group more therefore it was also a nonfunctional distractor (NF-D).
c. Item 46: distractors A, C, and E were all functional distractors (F-D) since they attracted more students from the lower group. However, distractor D was nonfunctional as it attracted less than 5% of students.

Discussion

According to Arikunto (2018), a good item is one that has a medium level of difficulty with an index starting from .31 to .70. In regard to the findings of the first question and objective of this research, medium items dominated the summative English test assessed to the 11th-grade students in MAN 2 Mataram in November 2022, with the percentage of 56%. This result is in line with several other studies that revealed the test items they had examined consisted of more items with moderate or medium levels (Maghfiroh, 2010; Ani, 2011; Risydah, 2014; Haryudin, 2015; Manfenrius et al., 2015; Pradanti et al., 2018; Maghfiroh, 2019; Jannah et al., 2021; and Marsevani, 2022). In contrast to that, Mehta and Mokhasi (2014), who analyzed the item difficulty level of 50 MCQs, found that the majority of the items (68%) were difficult. Moreover, in other research (Hartati & Yogi, 2019; Karlina, 2021), they found that easy items dominated the test with a percentage of 38% for the former and 82.5% for the latter.

The following crucial aspect analyzed in this research is item discriminating power. To determine the overall quality of test items, we cannot solely rely on analyzing the item difficulty level. For instance, an item with a medium level of difficulty does not always indicate that it is a perfect item, such as item 7 which was considered not very good although the level of difficulty was medium, it failed to discriminate between the high and low-achieving students. Therefore, this is where the next step to determine the quality of an item is needed. Based on the findings above, the English summative test analyzed in this research mostly consisted of satisfactory items (36% or 18 items). The second most items making up the test are good items (32% or 16 items). This means that less than 50% of the items were good at discriminating between the upper and lower group students, this is in line with other researchers (Toha, 2010; Risydah, 2014; Fajriah, 2016; Haryudin & Santosa, 2016; Pradanti et al., 2018; A. Maghfiroh, 2019). However, one distinction between this research and the others (Haryudin, 2015; Manfenrius et al., 2015) is that this research found one item with an excellent discriminating power whilst the other research revealed there were no items that met the criteria for an excellent item discriminating power. Next, this item also had 28% of poor items, this number is almost equal to that of the discriminating power analysis done by Quaigrain & Arhin (2017) which shows that 25% of the items were poor. Moreover, the finding of this research also showed that 1 item had a negative discriminating power which is in line with that found by Hartati & Yogi (2019).

Referring to the last findings of this research, the analysis of item distractors revealed that there were more functional distractors (54%) than nonfunctional (46%). On the contrary, Karlina (2021) obtained through the analysis that 78% (156) of the 200 distractors were nonfunctional. Other researchers (Risydah, 2014; Haryudin, 2015; Manfenrius et al., 2015; Pradanti et al., 2018) also found that there were more nonfunctional distractors within the tests they analyzed. It is important to remember that a functional distractor is one that is selected by at least 5% of the test takers and those in the lower group are expected to select it the most (Arikunto, 2018). If an item has more nonfunctional distractors, a revision is needed to improve the overall quality of the item.

CONCLUSION

The researcher concluded that in terms of the item difficulty level, the qualities of the English summative for the 11th-grade students in MAN 2 Mataram in the academic year 2022/2023 revealed that 19 items (38%) were easy, 28 items (56%) were medium, and 3 items (6%) were difficult. As for the item discriminating power, 1 item (2%) was considered very poor, 14 items (28%) were poor, 18 items (36%) were satisfactory, 16 items (32%) were good, and 1 item (2%) was excellent. Lastly, the effectiveness of distractors resulted in 108 functional distractors (54%) and 92 nonfunctional distractors (46%).

REFERENCES

Adom, D., Mensah, J.A., Dake, D.A. (2020). Test, measurement, and evaluation: Understanding and use of the concepts in education. *International Journal of Evaluation and Research in Education (IJERE).* doi: 10.11591/ijere.v9i1.20457

Ani, L. I. (2015). *An item analysis on the difficulty level of an english summative test for second grade of SMP Muhammadiyah 29 Cinangka-Sawangan Depok* (Bachelor thesis, Syarif Hidayatullah State Islamic University, Jakarta). Retrieved from: https://www.repository.uinjkt.ac.id/

Arikunto, S. (2018). *Dasar-dasar evaluasi pendidikan* (3rd ed.). Jakarta: Bumi Aksara.

Brown, H.D. (2003). *Language assessment: Principles and classroom practices*. New York: Pearson Longman.

Cizek, G.J. (2010). *An Introduction to formative assessment.* In Andrade, H.L., & Cizek, G.J. (2009). *Handbook of formative assessment*. New York: Routledge.

Ebel, R.L., and Frisbie, D.A. (1991) *Essentials of educational measurement* (5th Edition). Englewood Cliffs: Prentice-Hall Inc.

Fajriah, S. (2016). *An item analysis on discriminating power of English summative test* (Bachelor thesis, Syarif Hidayatullah State Islamic University, Jakarta). Retrieved from https://repository.uinjkt.ac.id/

Hartati, N., & Yogi, H. P. S. (2019). Item analysis for a better quality test. *Journal of English Language in Focus (ELIF), 2*(1), 59–70. Retrieved from https://jurnal.umj.ac.id/

Haryudin, A. (2015). Validity and reliability of English summative tests at Junior High School in West Bandung. *Jurnal Ilmiah UPT P2M STKIP Siliwangi, 2*(1), 77–90.

Haryudin, A., & Santosa, I. (2016). The analysis of discriminating power of English summative test. *Jurnal Ilmiah UPT P2M STKIP Siliwangi, 3*(2), 59–67.

Hikmawati, F. (2020). *Metodologi penelitian*. Depok: Rajawali Pers.

Hughes, A. (2003). *Testing for language teachers* (2nd eds). UK: Cambridge University Press.

Jannah, R., Hidayat, DN., Husna, N., Khasbani, M. (2021). An item analysis on multiple-choice questions: A case of a junior high school English try-out in Indonesia. *Jurnal Bahasa, Sastra dan Pengajarannya, 15* (1), 9-17. Retrieved from https://www.jurnalnasional.ump.ac.id/index.php/LEKSIKA/article/view/8768

Karlina, R. (2021). The item analysis of summative test at the eighth grade of SMPN 12 Lebong in the academic year of 2020/2021 (Bachelor thesis, State Institute for Islamic Studies, Bengkulu). Retrieved from https://repository.iainbengkulu.ac.id/

Kothari, C.R. (2004). *Research methodology: Methods and techniques* (2nd eds). Ansari Road, Daryaganji, New Delhi: New Age International.

Maghfiroh, F. (2010). *An item analysis of the difficulty level of an English summative test* (Bachelor thesis, Syarif Hidayatullah State Islamic University, Jakarta). http://repository.uinjkt.ac.id/dspace/bitstream/123456789/3860/1/FIFI%20MAGHFIROH-FITK.pdf

Maghfiroh, A. (2019). *An analysis on English midterm test for the second semester at the ninth grade of SMP TA'MIRUL ISLAM Surakarta* (Bachelor thesis), State Islamic Institute of Surakarta, Surakarta.

Manfenrius, A., Sutapa, G., & Wijaya, B. (2015). Item analysis on English summative test at the eighth grade Junior High Schools in Pontianak. *Jurnal Pendidikan Dan Pembelajaran Khatulistiwa, 4*(12), 1–10.

Marsevani, M. (2022). Item analysis of multiple-choice questions: An assessment of young learners. *English Review: Journal of English Education, 10*(2), 401-408. https://doi.org/10.25134/erjee.v10i2.6241

Mehta, G., & Mokhasi, V. (2014). Item analysis of multiple-choice questions-An assessment of the assessment tool. *International Journal of Health Science and Research, 4*(7), 197-202.

Pradanti, S. I., Martono, M., & Sarosa, T. (2018). An item analysis of English summative test for the first semester of the third grade Junior High School students in Surakarta. *English Education Journal, 6*(3), 312–318. https://doi.org/10.20961/eed.v6i3.35891

Quaigrain, K., & Arhin, A.K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education, 4*(1), 1301013

Risydah, Y. (2014). *An analysis of test items of the English first-term test of the seventh-grade students of SMP Muhammadiyah 10 Yogyakarta in the Academic Year of 2013/2014* (Bachelor thesis, Universitas Negeri Yogyakarta). https://eprints.uny.ac.id/18498/1/Yunita%20Risydah%2006202244051.pdf

Shepard, L. A. (2019). Classroom Assessment to Support Teaching and Learning. *The ANNALS of the American Academy of Political and Social Science, 683* (1), 183–200. https://doi.org/10.1177/0002716219843818

Toha, H. (2010). *An item analysis of English summative test for the first year of Junior High School* (Bachelor thesis, Syarif Hidayatullah State Islamic University, Jakarta). Retrieved from https://repository.uinjkt.ac.id/

Woods, N. (2015). Formative assessment and self-regulated learning. *The Journal of Education*. Retrieved from https://thejournalofeducation.wordpress.com