

Item Analysis of an English Achievement Test Using Rasch Model and Classical Test Theory

Rezkiaturahmi^{a,1*}, Nur Hidayanto Pancoro Setyo Putro^{a,2}, Guldana Begimbetova^{b,3}, Oscar Ndayizeye^{c,4}, Muthmainnah^{d,5}, Dinda Khaira Latifa^{a,6}

^a Universitas Negeri Yogyakarta, Yogyakarta, Indonesia

^b Abai Kazakh Pedagogical University, Khazakhstan

^c Hebei Foreign Studies University, China

^d Universitas Muhammadiyah Kendari, Kendari, Indonesia

¹ rezkilaturahmi98@gmail.com; ² nur_hidayanto@uny.ac.id; ³ begimbetovaguldana227@gmail.com;

⁴ oscarndayiz@gmail.com; ⁵ dindakhaira.2024@student.uny.ac.id

* Corresponding Author

ARTICLE INFO

Article history

Received August 04, 2026

Revised September 08, 2026

Accepted October 11, 2026

Keywords

English Test;

Classical Test Theory;

IRT (Rasch Model)

©2025 The Author.

This is an open-access article under the CC-BY license.



ABSTRACT

This study aimed to evaluate the quality of English test items developed by the English Teachers Working Group (MGMP) in Muna District, Indonesia. The study specifically examined the psychometric characteristics of the items in terms of item difficulty, discrimination, distractor effectiveness, and item fit using Classical Test Theory (CTT) and Item Response Theory (IRT). This study employed a quantitative psychometric approach. The data consisted of responses to an English achievement test developed and administered by English teachers participating in the MGMP in Muna District. Item quality was initially examined using Classical Test Theory based on item difficulty, discrimination index, and distractor effectiveness. Further analysis was conducted using the Rasch model of Item Response Theory with the *Quest* program to evaluate item difficulty and item fit based on INFIT Mean Square (MNSQ) statistics. The results of both approaches were used to determine whether individual items were acceptable, required revision, or should be rejected. The analysis revealed variation in the psychometric quality of the test items. The CTT analysis showed differences in item difficulty, discrimination, and distractor effectiveness, while the Rasch analysis revealed variation in item difficulty and item fit. The combined findings indicated that some items demonstrated acceptable psychometric characteristics, whereas others required revision. In addition, several items could not be evaluated using *Quest*. These findings indicate that although more than half of the items demonstrated acceptable psychometric characteristics, a substantial proportion still requires improvement before being used as a high-quality English assessment.

1. Introduction

In the 21st century, evaluating students' abilities in educational settings is achieved through assessment. These assessment skills are essential for both teachers and students [1]. They serve as a primary standard for gauging student abilities. The main goal of assessment is to review the teacher's approach to teaching and learning, providing feedback that helps refine teaching methods to meet educational objectives [2][3][4][5].

Essentially, assessments are crucial in gauging the effectiveness of learning and determining the competencies students acquire after class participation. They are used to evaluate the educational process and are implemented after conducting tests or assessments on students. However, it's

necessary for teachers to use various approaches to evaluate student achievements and gather information on learning outcomes [6][7]. Therefore, assessment is an ongoing, comprehensive process used to collect information on student performance and determine their success in the learning process.

In addition, [8][9] assessment for learning is a crucial process in education, centered around gathering information about students' knowledge, skills, and understanding. This process aids teachers in shaping their instructional strategies [10][11]. It can be integrated continuously into the curriculum or implemented at significant junctures, such as end-of year examinations. The primary aim of assessment for learning is not merely to test student knowledge but to enhance the effectiveness of teaching. It helps educators refine their lesson plans, improve learning outcomes, and pinpoint both the strengths and areas for improvement in their students. This type of assessment offers students various ways to showcase their learning, including written assignments and practical demonstrations. Furthermore, also [12][13] assessment also plays a vital role in the educational process, serving as a key tool for teachers to gauge the success of their teaching strategies and to determine how well students are grasping the material taught. In educational settings, there are primarily two types of assessments namely summative and formative. Therefore, formative also known as Assessment for Learning (*AfL*) is a more continuous and interactive approach. Because, [11][14] it is designed to provide constructive feedback during the learning process, helping students to improve their performance as they are still learning. Thus, this type of assessment encourages students to take a more active role in their learning, fostering an understanding of what constitutes successful work and promoting independence through peer and self-assessment.

The Item Response Theory (IRT) emerged as a modern alternative to classical test theory, addressing its limitations [15][16]. It was developed by experts in psychology and education to create more valid and reliable measurements. IRT is characterized by features such as item characteristics being independent of the students, participants' ability scores being dependent on the test and the ability to provide appropriate measures for each ability score without needing parallel tests for reliability. IRT comprises various models based on the number of parameters the use: the one parameter logistic model (1PL), two parameter logistic model (2PL), three parameter logistic model (3PL), and four parameter logistic model (4PL) [17][18][19][20][21][22]. The focus of this research is on the 1PL model, also known as the Rasch model. In addition, Georg Rasch a pioneer in this field, developed the 1PL model. His groundbreaking work in the 1950s involved analyzing elementary school students' exam result, leading to the insight that the probability of a student correctly answering an item is equal when the student's ability matches the item's difficulty level. Hence, this principle forms the core of the Rasch model, simplifying and refining the measurement process.

In addition, [15][23][24][25] classical test analysis uses traditional items and sample-dependent statistics. This includes item difficulty and item discrimination estimates, distractor analysis, item correlations and various related statistics. Unfortunately, most psychometric analyses focus on assessing examinees at the test score level rather than at the item level. Classical test analysis also generally includes measures of score reliability, item difficulty and instrument discrimination. Item analysis is a set of statistical procedures that focus on selecting items to maximize score reliability. The main classical analysis statistics are item difficulty, item discrimination and instrument reliability. Item difficulty in classical test theory (CTT) is one of the characteristics of an item. Item characteristics are the proportion of examinees who take the test and those who answer the item correctly [20][26][27][28][29]. The greater the percentage of items answered correctly by examinees, the easier the item is. The higher the difficulty value, the easier the item is to understand. The item difficulty index is calculated by dividing the number of examinees who answer the item correctly by the total number of examinees who answer the item.

One of the important parameters in classical test theory analysis is the evaluation of distractor function. [30] stated that the effectiveness of distractors is measured based on the number of test takers who choose them. [31] identified two main criteria for assessing distractor function, namely ability to choose and discrimination power. First, a distractor is considered functional if it is chosen by at least some test takers, with a minimum recommendation of 5% as the limit for maintaining it. If no participants choose the distractor, then the distractor is considered ineffective and needs to be replaced.

Second, a distractor is considered to have good discrimination ability if it is chosen more often by low-ability participants compared to high-ability participants. This shows that effective distractors have a negative correlation with the total test score. Conversely, distractors with a positive or zero correlation to the total test score are considered ineffective and can reduce the overall quality of the test items.

In Indonesia, teachers often use formative tests to assess students at the middle of the learning process. A test as a set of tools designed to measure an individual's proficiency within specific criteria [32]. In addition, a test to assess students' abilities through a series of questions within a given timeframe. Tests are thus seen as instrumental in evaluating students' competencies and tracking their progress. These tests are key in uncovering students' competencies after a school term or course. Consequently, it is imperative for teachers to construct effective tests. Writing tests for evaluation purposes has been a longstanding challenge for educators [33]. In addition, the main difficulty lies in crafting questions that are both effective and fair [34]. These questions should ideally balance the level of difficulty with a discrimination index, which measures the ability of a question to distinguish between students who have mastered the content and those who have not.

Unfortunately, some teachers are less concerned about constructing a good assessment. There are still many tests that only assess the lower levels of thinking, and the questions have not been tested so the quality is not known, and it is feared. [35] conducted items quality study using classical analysis to uncover the distractor effectiveness and validity of the items using classical test theory and found that the distractor effectiveness was functioning 72.5% while those that did not function were 27.5%. In addition, [30] conducting item analysis of English test and the results show that most of the questions are included in the medium category (92.5%) and have excellent differentiating power (97.5%), with 80% effective distractors. Considering the problem above, there is a need for studies and trials on the end of the semester exam to know the quality of the item questions. The research used the Quest program with the classical test theory and IRT (Rasch model) to gain the items quality. This study will be developed from the previous studies. The difference between the recent study and the previous study lies in the subject of the study which is the previous study focus the quality of the English test that is questions given to students in final examination and to find out the item's quality based on classical test theory and IRT (Rasch model).

For that reason, this study is very important to do because exam questions are a vital instrument in an assessment, especially to find out student abilities or competencies that have been achieved by students after studying for one semester, so it is necessary to know the quality based on classical test theory and IRT (Rasch model). Thus, this analysis result can provide comprehensive information regarding the quality of the level of questions that is used in the examination. In addition, the results of this study can give a contribution to schools, especially to the teachers as a reference to improve the quality of the test that is being tested in exam to improve the quality of graduates in accordance with the demand of curriculum of 2013.

The present study addresses this gap by examining an English achievement test developed by English teachers participating in the MGMP in Muna District, Southeast Sulawesi, Indonesia. An English test developed through the MGMP deserves systematic psychometric evaluation because it is not merely an individual classroom assessment but a collaboratively developed assessment that may be used across participating teachers or schools. The items are expected to represent common learning objectives and provide comparable information about students' English achievement. However, collaborative test development does not automatically guarantee that all items function appropriately. Differences in teachers' experience, item-writing practices, interpretation of learning indicators, and construction of distractors may result in items with different levels of difficulty and discrimination. Without empirical item analysis, problematic items may remain undetected and potentially affect the accuracy of decisions based on test results. Therefore, analyzing MGMP-developed English tests is important to determine whether the items function as intended and to provide evidence for their continued use, revision, or replacement.

Despite the important role of MGMP in supporting collaborative assessment practices, empirical studies examining the psychometric quality of MGMP-developed English examinations remain

limited, particularly at the district level. Previous studies have generally focused on individual classroom tests or have applied a single psychometric framework, providing only partial evidence of item functioning. In this study, Classical Test Theory (CTT) and the Rasch model are used as complementary approaches. CTT provides practical information about item difficulty, discrimination, and distractor effectiveness, while Rasch analysis provides model-based evidence concerning item difficulty and item fit. Combining these approaches allows the quality of MGMP-developed English test items to be examined from both empirical item statistics and model-based measurement perspectives. Accordingly, this study aims to provide a more comprehensive evaluation of the English achievement test developed by the MGMP in Muna District.

Based on this rationale, the study seeks to answer three related questions namely (a) how do the English test items developed by the MGMP in Muna District perform in terms of item difficulty, item discrimination, and distractor effectiveness based on CTT?, (b) how do the items perform in terms of item difficulty and item fit based on the Rasch model, and (c) how can the findings from CTT and Rasch analysis be integrated to determine which items are acceptable, require revision, or require further evaluation? By addressing these questions, the study is expected to provide empirical evidence that can support English teachers and the MGMP in improving the quality of collaboratively developed assessments and strengthening the basis for future test development.

Thus, the contribution of this research is to provide empirical evidence on the psychometric quality of an English examination developed collaboratively by the English Teachers Working Group (MGMP) in Muna District. By combining Classical Test Theory (CTT) and the Rasch model, this study provides complementary information on item difficulty, discrimination, distractor effectiveness, and item fit. The findings provide a practical basis for the MGMP English teachers to identify items that can be retained, revised, or further examined and to support the improvement of future English assessment development.

2. Method

This study employed a descriptive quantitative research design to investigate the psychometric quality of English test items administered to Grade 12 students in senior high schools in Muna District, Southeast Sulawesi, Indonesia, during the 2023/2024 academic year. The study focused on an English examination developed and used by the English Teachers Working Group (Musyawarah Guru Mata Pelajaran/MGMP) in Muna District. The examination was analyzed to determine the extent to which the test items met appropriate psychometric criteria and were suitable for measuring students' English achievement.

The population of this study consisted of Grade 12 students who participated in the English examination in Muna District during the 2023/2024 academic year. The data were collected from five senior high schools participating in the examination. The schools included five senior high schools in Muna District. These schools represented different school contexts within Muna District and provided a sufficiently large number of student responses for conducting psychometric analysis. Initially, the total number of registered students across the participating schools was 1,131 students. However, 14 students did not participate in the English examination. Therefore, their responses were not available for analysis. After excluding these students, the final dataset consisted of 1,117 students who completed the English examination. Thus, the analysis was based on 1,117 student response patterns to the English test items. The participating schools were selected because the English test developed by the English Teachers Working Group (MGMP) in Muna District was newly developed and was piloted in these schools. Therefore, the selection of the schools was based on their participation in the implementation and piloting of the MGMP-developed English test rather than random sampling. All eligible Grade 12 students who participated in the examination in the selected schools were included in the study.

The instrument analyzed in this study was the English examination developed by the English Teachers Working Group (MGMP) in Muna District for Grade 12 students in the 2023/2024 academic year. The test consisted of 45 items covering two forms of assessment. The first section consisted of

40 multiple-choice items, while the second section consisted of five essay items. In addition, the 40 multiple-choice items were developed based on the curriculum and learning objectives for senior high school students using the 2013 Curriculum (Kurikulum 2013). The items were designed to represent the English learning competencies and curriculum indicators relevant to Grade 12 students. The test blueprint specified the learning indicators, content areas, and cognitive levels represented by the items. The multiple-choice items were the primary focus of the item level psychometric analysis because their response structure allowed the researcher to examine item difficulty, discrimination, and distractor functioning. In particular, the multiple-choice items provided information about the extent to which each item could differentiate between students with different levels of English ability and whether the alternative options functioned effectively. The test was developed within the MGMP context, meaning that the items represented the collective assessment practices of English teachers involved in the development of the examination. Consequently, analyzing the quality of these items is important not only for evaluating the examination but also for providing feedback to teachers regarding the development of future English assessments.

Furthermore, the interpretation of the item analysis was based on predetermined psychometric criteria. For CTT, item difficulty was interpreted according to the proportion of students who answered each item correctly, with higher values indicating easier items. Item discrimination was interpreted based on the discrimination index used in the analysis, while distractor effectiveness was evaluated based on the proportion of students selecting each distractor and its relationship with overall test performance. For the Rasch analysis, item fit was evaluated using the INFIT MNSQ statistic based on the criterion specified for the analysis.

The item analysis was conducted using the *QUEST* program. *QUEST* is a computer-based program that can be used to conduct item analysis and Rasch-based analysis of educational assessment data. In this study, the program was used to obtain empirical information regarding the characteristics and quality of the English test items. The analysis proceeded by entering the prepared student response data into the *QUEST* program. The program then generated statistical outputs that were interpreted to evaluate the quality of the test items. The outputs used in this study included information relevant to item difficulty and item fit, particularly the INFIT MNSQ statistics. The results were subsequently interpreted together with the findings obtained from the Classical Test Theory analysis. The use of the *QUEST* program also enabled the researcher to evaluate the consistency of individual items with the Rasch measurement model. Therefore, the analysis did not merely identify whether an item was easy or difficult but also examined whether the item behaved in accordance with the underlying measurement model.

The results of both approaches were then compared to identify items that could be classified as acceptable, items requiring revision, and items that were not suitable for continued use. This procedure allowed the researcher to make a more comprehensive judgment about item quality rather than relying on a single statistical indicator.

3. Results and Discussion

3.1. The Item Quality based on Classical Test Theory

The result of analysis using the classical test theory have been tested by *Quest* program that indicate the characteristics of quality of the item questions of English test items developed by English teacher working group MGMP in Muna district. In addition, the results of the items using the classical test theory are the difficulty level (Table 1), discrimination index and the distractor effectiveness.

The difficulty distribution suggests that the test was not sufficiently balanced across the ability range of the examinees. The strong concentration of very easy items may have limited the test's capacity to challenge higher-performing students and provide adequate measurement across different levels of English proficiency. A more balanced distribution of item difficulty would allow the test to capture students' ability differences more effectively.

Table 1. Difficulty Level

Category	Items	The Number of Items	Percent
Very Difficult <30%	21 and 35	2	5%
Difficult 30-40%	23	1	2.5%
Moderate 41-84%	5, 8, 9, 11, 13, 14, 15, 24, 28, 31 and 34	11	27.5%
Easy 85-90%	-	-	0%
Very Easy >90	1, 2, 3, 4, 6, 7, 10, 12, 16, 17, 18, 19, 20, 22, 25, 26, 27, 29, 30, 32, 33, 36, 37, 38, 39 and 40	26	65%

The discrimination results (Table 2) indicate that a considerable proportion of the test items still had limited capacity to distinguish students according to their level of achievement. The presence of items with weak or negative discrimination characteristics warrants particular attention because such items may not contribute positively to the measurement of students' English ability. These items should therefore be reviewed in relation to their content, wording, answer keys, and alignment with the intended learning outcomes before being reused.

Table 2. Discrimination Index

Category	Items	The Number of Items	Percent
Good item >0.40	6, 9, 23, 32, 35 and 38	6	15%
No revision required 0.30-0.39	1, 2, 3, 4, 7, 10, 14, 15, 19, 24, 25, 30, 31, 33, 37 and 39	16	40%
Items require revision 0.20-0.29	16, 17, 18, 20, 22, 26 and 27	7	17.5%
Items must be eliminated <0.19	5, 8, 11, 12, 13, 21, 28, 29, 34, 36 and 40	11	27.5%

The distractor analysis (Table 3) indicates that the functioning of the alternative options was not evenly distributed across the test. This suggests that some distractors were more attractive to examinees than others, while several alternatives may have had limited effectiveness. From an item-development perspective, distractors that are rarely selected should be reviewed because they may be too implausible, obviously incorrect, or insufficiently related to the item stem. Improving distractor quality could increase the discriminatory potential of the multiple-choice items. In other words, Based on the results of the overall analysis of the items using *Quest* program with the classical test theory approach in Muna district high school, the results show that the difficulty level is 2 (5%) items in the very difficult category, 1 (2.5%) item question in the difficult category, 11 (27.5%) item questions in the moderate category and 26 (65%) item questions in the very easy category. In addition, in terms of the results of the discrimination index, it shows that 6 (15%) item questions are in the good item category, 16 (40%) item questions are in the no revision required category, 7 (17.5%) item questions are in the items require revision category and 11 (27.5%) item questions category items must be eliminated. On the other hand, based on the results of the distractor effectiveness, it shows that option A works on the 5 item questions with a percentage of 12.5%. In addition, option B works on 3 item questions with a percentage of 20%, option C works on 5 item questions with a percentage of 12.5%, option D works on 6 item questions with a percentage of 15% and option E works on 2 item questions with a percentage of 5%.

Table 3. The Distractor Effectiveness

Options	Item Questions	The Number of Items	Percent
A	8, 9, 31, 35 and 38	5	12.5%
B	14, 23 and 35	3	20%
C	5, 11, 21, 28 and 34	5	12.5%
D	9, 11, 13, 15, 24 and 34	6	15%
E	21 and 23	2	5%

3.2. The Items Quality based on IRT (Rasch Model)

The results of analysis (Table 4 and Table 5) using the item response theory approach that have been tested by using Quest program produce item statistics that indicate the characteristics or quality of the item questions of English test developed by English teacher working group MGMP in Muna district. Besides, the results of characteristics of the items using Rasch model approach are the difficulty level and the quality of the items according to the criteria of item compatibility with Rasch model (Infit MNSQ).

Table 4. The Result of Difficulty Level based on IRT (Rasch Model)

Category	Item	The Number of Items	Percent
Not analyzed	12, 29, 36 and 40	4	10%
Very Difficult $b > 2$	5, 8, 9, 11, 13, 14, 15, 21, 23, 24, 28, 31, 34 and 35	14	35%
Difficult $1 < b \leq 2$	32 and 38	2	5%
Moderate $-1 < b \leq 1$	6, 20 and 25	3	7.5%
Easy $-1 > b \geq -2$	1, 2, 3, 4, 7, 10, 19, 30, 33, 37 and 39.	11	27.5%
Very Easy $b < -2$	16, 17, 18, 22, 26 and 27	6	15%

Table 5. The Result of INFIT MNSQ based on Rasch Model

Category	Items	Number of Items	Percent
Not analyzed	12, 29, 36, 40	4	10%
Not compatible with Rasch model $1.30 > \text{INFIT MNSQ}$	-	-	-
Compatible with Rasch model $0.77 < \text{INFIT MNSQ}$	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 30, 31, 32, 33, 34, 37, 38 and 39	35	87.5%
Not compatible with Rasch model $\text{INFIT MNSQ} < 0.70$	35	1	2.5%

Based on the Table 4 and Table 5, it shows that the 4 questions cannot be analyzed, or the difficulty value cannot be determined. In addition, 14 questions were found in the very difficult category with a percentage of 35%. On the other hand, difficult questions are found in question numbers 32 and 38 with a percentage of 5%. Meanwhile, the questions that were found to be moderate with a 7.5% percentage. In addition, 11 questions were found in the easy category with a percentage of 27.5% and 6 questions were found in the very easy category with a percentage of 15%. Hence, the presentation of very difficult questions dominated the English test with a total presentation of 35%. On the other hand, regarding the INFIT MNSQ result shows that there are 35 (87.5%) items that compatible with the Rasch model based on the Infit MNSQ value of the Quest program. Meanwhile, there is a question number 35 (2.5%) that is not compatible with the Rasch model. Therefore, the question that compatible with Rasch model dominated in the English test items.

3.3. Item Qualities based on Classical Test Theory (CTT) and IRT (Rasch Model)

The results of the interpretation of item quality on the English test items was based on the results of the analysis of the level of difficulty, discrimination index, distractor effectiveness by using classical test theory and the level of difficulty, the value of the Infit MNSQ based on IRT (Rasch model), then draw conclusions from the quality of the item. Fig. 1 is the interpretation of items quality.

Based on the table above, it shows that there 22 (55%) items that were categorized into accepted items. 14 (35%) items that were categorized into revised questions. In addition, 4 questions cannot be analyzed by *Quest* program. In short, the item questions that were accepted dominated in the English test items developed by English teachers working group MGMP in Muna district. Thus, for items that have very good and good categories, they deserve to be included in the question bank to be used as an evaluation for the next exam. Meanwhile, for items that have moderate and weak categories, it is best to revise or review again and adjust to the deficiencies of each item in terms of difficulty level, discrimination index and distractor effectiveness. Therefore, questions that have a bad and very bad

category should not be used and replaced with new questions because these item questions are very unfit to be used as further evaluation instruments.

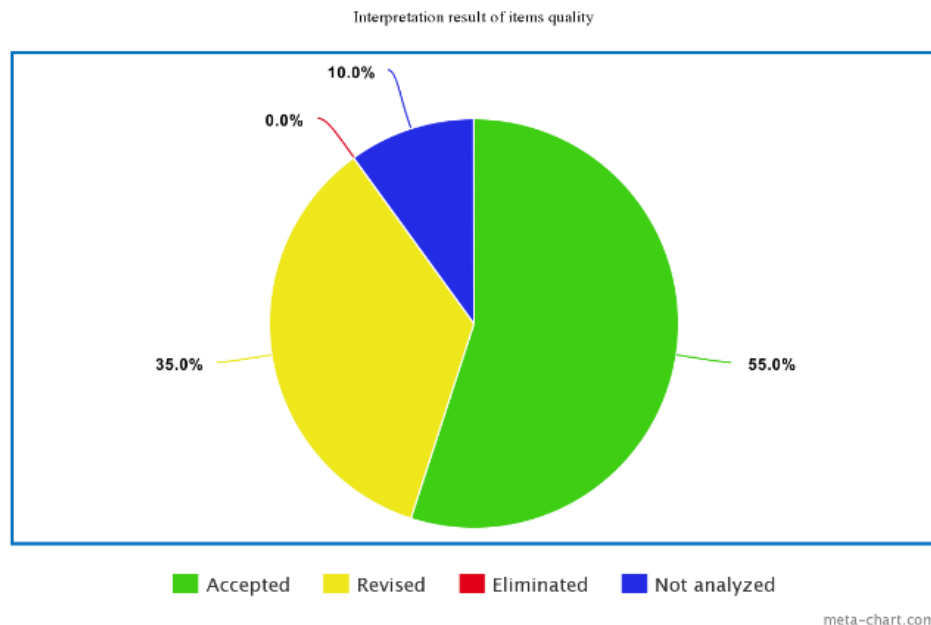


Fig. 1. The Percentage of Item Qualities

The design of the item analysis in English test items developed by English teacher working group MGMP at 2023/2024 academic year was carried out using *Quest* program with classical test theory and IRT (Rasch model). Analysis using *Quest* program produces item parameters in the form of level of difficulty, discrimination index, and distractor effectiveness based on classical test theory while difficulty level and infit MNSQ based on IRT (Rasch model). Therefore, the statistical instrument test is an estimate of each item against the respondent which shows the quality of the test instrument. Moreover, the results of item quality using *Quest* program by looking at the level of difficulty, discrimination index, and Infit MNSQ showed that 14 items need to be revised. However, because the teacher-made test is mainly an English test items developed by the English teacher working group MGMP, at least it does not require a trial run, the test is standardized, so that the quality of the items is forgotten. Additionally, the assessment process carried out by the teacher was to measure student success [6], [7], [11]. The assessment is carried out through the instrument. Assessment of learning is inseparable from the role of the teacher [33], [36]. It means that, the teacher has the full right to assess students. However, teachers have a strong foundation when facing assessments related to the assessment process in the 2013 curriculum. In addition, several foundations for establishing assessments in the 2013 curriculum, namely attitudes of competence, knowledge, and balance of skills. Assessment of students will consider the standards that have been applied to the 2013 curriculum. The scope of the 2013 assessment includes learning materials, competencies, subjects, content competencies, program competencies and processes [37]. Therefore, the scope of assessment in the 2013 curriculum does not only focus on one but other competencies are taken into consideration. That's why, the instrument used must have national standards in quality to be able to find out the weaknesses of education. Therefore, the quality of the items greatly influences the assessment used as evaluation material in education.

In addition, many teachers overlook the quality of the items by not doing a trial run on students. That's why, the information about the results of students' learning as one of the most important aspects, administering tests in teaching language also provides information about the accuracy of identification of purposes, suitability of materials teaching, the suitability and effectiveness of teaching methods, the adequacy of the exercises given, and student learning difficulties [30]. This is agreed because in the assessment process, the teacher will carry out an objective assessment that focuses on certain aspects. Therefore, a test is us as a measurement and an assessment tool because

the test results obtained can be in the form of student characteristics and student cognitive results [32]. In general, the quality of the items is important as evaluation material to increase accuracy in helping goals, aligning subject matter, aligning the effectiveness of teaching methods, the adequacy of the exercises given and learning difficulties. All that information can be obtained since observation on the level of student' success as reflected in the values achieved, and studies on the level and type of errors made by students. By given so, the importance of the function of the test, that's how necessary it is to try to use the test in teaching is really a good and reliable tool.

4. Conclusion

This study evaluated the psychometric quality of 40 multiple choice English test items developed by the English Teachers Working Group (MGMP) in Muna District using Classical Test Theory (CTT) and the Rasch model through the *Quest* program. The findings indicate that the items varied in terms of difficulty, discrimination, distractor effectiveness, and Rasch item fit. Based on the combined results, 22 items were classified as acceptable, 14 items required revision, and 4 items could not be analyzed using *Quest*. These findings indicate that although a substantial proportion of the items demonstrated acceptable psychometric characteristics, several items still require further review before the test is used for subsequent assessment.

The findings highlight the importance of conducting systematic item analysis in the development of MGMP-based English assessments. The 14 items identified as requiring revision should be reviewed particularly regarding their item characteristics and construction, while the four items that could not be analyzed using *Quest* require further examination to determine the reasons for their exclusion from the Rasch analysis. The MGMP teachers are therefore recommended to use the results of the CTT and Rasch analyses as empirical evidence for revising, replacing, or retaining test items in future examinations. Such systematic review can support the development of more consistent and appropriate English assessments for Grade 12 students. Therefore, the quality of the items affects student assessment so that it requires an increase in the preparation of questions for teachers in terms of cognitive level and the quality of the items in terms of the value of the difficulty level, the discriminating power index, the effectiveness of the distractor based on classical test theory while the level of difficulty and INFIT MNSQ based on IRT (Rasch model).

Supplementary Materials: No supplementary materials are available for this study.

Author Contribution: Conceptualization, methodology, data collection, data analysis, writing—original draft preparation, and writing—review and editing were carried out by the authors. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The APC was not funded by any external organization.

Acknowledgment: The authors would like to thank the English Teachers Working Group (MGMP) in Muna District, the principals and English teachers of the participating schools, and the students who participated in the English examination. The authors also acknowledge the support provided by the participating schools during the data collection process.

Conflicts of Interest: The authors declare no conflict of interest.

References

- [1] Z. Edger-lacoursière *et al.*, "The MacHAND performance assessment (MPA): Development and psychometric testing of the short English version (MPA-S) for the traumatic hand injury population," *J. Hand Ther.*, vol. 39, no. 1, pp. 156–167, 2026, <https://doi.org/10.1016/j.jht.2025.04.008>.
- [2] M. Kawata, M. Tsuruta-hamamura, and H. Hasegawa, "Assessment of speech transmission index and reverberation time in standardized English as a foreign language test rooms," *Appl. Acoust.*, vol. 202, p. 109093, 2022, <https://doi.org/10.1016/j.apacoust.2022.109093>.

- [3] M. Paquot *et al.*, “The Varieties of English for Specific Purposes dAtabase (VESPA): Towards a multi-L1 and multi-register learner corpus of disciplinary writing,” vol. 10, no. 2, pp. 1–15, 2022, <https://doi.org/10.32714/ricl.10.02.02>.
- [4] V. Gamboa, M. P. Paixão, and S. Neves de Jesus, “Internship quality predicts career exploration of high school students,” *J. Vocat. Behav.*, vol. 83, no. 1, pp. 78–87, 2013, <https://doi.org/10.1016/j.jvb.2013.02.009>.
- [5] V. Gamboa, M. P. Paixão, and S. N. de Jesus, “Vocational profiles and internship quality among Portuguese VET students,” *Int. J. Educ. Vocat. Guid.*, vol. 14, no. 2, pp. 221–244, 2014, <https://doi.org/10.1007/s10775-014-9268-0>.
- [6] R. Handapangoda and J. Kormos, “Scene behind the unseen: Special adjustments for visually impaired students in L2 English reading tests,” *System*, vol. 140, p. 104049, 2026, <https://doi.org/10.1016/j.system.2026.104049>.
- [7] Q. Si, “Deep learning-based English speech recognition and pronunciation assessment system,” *Syst. Soft Comput.*, vol. 9, p. 200570, 2026, <https://doi.org/10.1016/j.sasc.2026.200570>.
- [8] S. Wang, S. Min, H. Wang, X. Gao, and N. Ding, “Computers and Education : Artificial Intelligence Standardized assessment of LLM English proficiency,” *Comput. Educ. Artif. Intell.*, vol. 11, p. 100656, 2026, <https://doi.org/10.1016/j.caeai.2026.100656>.
- [9] Z. Liu, B. Igried, M. Aljaidi, A. Abdullah, A. Fayez, and M. Khishe, “A pedagogically-guided wavelet-enhanced autoencoder for intelligent assessment of english language teaching,” *J. Eng. Res.*, 2026, <https://doi.org/10.1016/j.jer.2026.08.014>.
- [10] L. Qi and P. Shi, “Study on multi-dimensional assessment model of English reading ability based on SHAP and LIME interpretation,” *Syst. Soft Comput.*, vol. 9, p. 200591, 2026, <https://doi.org/10.1016/j.sasc.2026.200591>.
- [11] M. P. Saadi *et al.*, “Assessment of Interrater Reliability in Point-of-Care Ultrasound for Assessing Congestion in Cardiovascular Intensive Care,” *ABC Imagem Cardiovascular*, 38, no. 2, 2025, <https://doi.org/10.36660/abcimg.20250022i>.
- [12] S. Park, S. B. Piasta, and P. M. Sayer, “Early Childhood Research Quarterly An initial yet rigorous test of multisensory alphabet instruction for english monolingual and emergent bilingual children,” *Early Child. Res. Q.*, vol. 72, pp. 328–338, 2025, <https://doi.org/10.1016/j.ecresq.2025.04.008>.
- [13] W. Zhang *et al.*, “Evaluation of the effectiveness of using flipped classroom in puncture skills teaching,” *BMC Med. Educ.*, vol. 24, no. 1, pp. 1–11, 2024, <https://doi.org/10.1186/s12909-024-05132-8>.
- [14] Y. Kuo, S. I. C. Wang, E. Zhi, and F. Liu, “Peer assessment using one-to-one tablets to improve the English oral presentation of fourth-grade students in Taiwan,” *Comput. Educ. Open*, vol. 8, p. 100262, 2025, <https://doi.org/10.1016/j.caeo.2025.100262>.
- [15] Y. Liu *et al.*, “Development of the Short Form of the Specific Module of QLICD-BPH (V2.0) for Assessing Quality of Life with Benign Prostatic Hyperplasia by Classical Test Theory, Item Response Theor, and Generalizability Theory,” *Eur. Urol. Open Sci.*, vol. 91, pp. 32–40, 2026, <https://doi.org/10.1016/j.euros.2026.05.022>.
- [16] M. Adler, J. Skånberg, U. Brodin, and F. Lenhard, “Journal of Affective Disorders Reports An item response theory analysis of PHQ-9,” *J. Affect. Disord. Reports*, vol. 26, p. 101150, 2026, <https://doi.org/10.1016/j.jadr.2026.101150>.
- [17] J. A. Marmara and D. Zarate, “An item response theory examination of the functionality appreciation scale,” *Body Image*, vol. 58, p. 102133, 2026, <https://doi.org/10.1016/j.bodyim.2026.102133>.
- [18] M. Paz and L. Ezquerra, “Measuring overconfidence: An item response theory approach,” *J. Math. Psychol.*, vol. 130, p. 103003, 2026, <https://doi.org/10.1016/j.jmp.2026.103003>.
- [19] E. Yusron, H. Retnawati, and I. Rafi, “Bagaimana hasil penyetaraan paket tes USBN pada mata pelajaran matematika dengan teori respon butir?,” *J. Ris. Pendidik. Mat.*, vol. 7, no. 1, pp. 1–12, 2020, <https://doi.org/10.21831/jrpm.v7i1.31221>.

- [20] M. Molaei, P. Fadakar, and M. Shafieetabar, "Psychometric properties and factor structure of the cognitive distortions questionnaire (CD Quest)," *Psychiatry Res. Commun.*, vol. 5, no. 4, p. 100241, 2025, <https://doi.org/10.1016/j.psychom.2025.100241>.
- [21] S. A. Livingston, "Equating Test Scores (without IRT)," *ETS Rep.*, pp. 1–80, 2004, [Online]. Available: papers3://publication/uuid/753FF7E7-6A9F-4F37-99FA-5FC927542973
- [22] O. N. Bakytbekovich *et al.*, "Distractor Analysis in Multiple-Choice Items Using the Rasch Model," *Int. J. Lang. Test.*, vol. 13, pp. 69–78, 2023, https://www.ijlt.ir/article_169049.html.
- [23] J. R. Rico-juan and M. Arevalillo-herráez, "D-CTT : A deterministic and scalable heuristic for real-time item calibration in adaptive learning systems," *Expert Syst. Appl.*, vol. 333, p. 134180, 2027, <https://doi.org/10.1016/j.eswa.2026.134180>.
- [24] Y. Altın, B. Ozkurt, and H. F. Küçükibis, "Acta Psychologica Examination of the psychometric properties of physical activity enjoyment scale (PACES) based on classical test theory and item response theory," vol. 261, 2025, <https://doi.org/10.1016/j.actpsy.2025.105855>.
- [25] L. Kun, Y. Yi, Y. Bing, M. Rui, and G. Lei, "International Journal of Nursing Studies Advances Validation of the short index of job satisfaction in Chinese nurses: classical test theory and item response theory," *Int. J. Nurs. Stud. Adv.*, vol. 8, p. 100321, 2025, <https://doi.org/10.1016/j.ijnsa.2025.100321>.
- [26] Y. Dai, C. Chen, Z. Lin, J. Qin, and Y. Guo, "Computers & Education Development and validation of a scenario-based multiple-choice test of AI ethical competence (AIEC) using item response theory," *Comput. Educ.*, vol. 251, p. 105641, 2026, <https://doi.org/10.1016/j.compedu.2026.105641>.
- [27] P. Chen, Y. Li, M. Xiong, and Y. Zhang, "Acta Psychologica Psychometric evaluation of the behavioral inhibition/activation system scales in older adults in Mainland China: A classical test theory and item response theory approach," vol. 258, no. 783, 2025, <https://doi.org/10.1016/j.actpsy.2025.105280>.
- [28] N. S. Thamrin *et al.*, "Eclectic Approach : A Search for More Effective Assessment Ways to Meet EFL Assessment Principles," vol. 10, no. 2, pp. 668–685, 2023, <https://doi.org/10.24815/siele.v10i2.26682>.
- [29] L. Huang *et al.*, "The Journal of nutrition, health and aging Development validation of a ultra-brief screening tool for sarcopenia in Chinese community-dwelling older adults," *J. Nutr. Heal. aging*, vol. 30, no. 9, p. 100928, 2026, <https://doi.org/10.1016/j.jnha.2026.100928>.
- [30] A. Maharani and N. Putro, "Item Analysis of English Final Semester Test," *Indones. J. EFL Linguist.*, vol. 5, p. 491, Dec. 2020, <https://doi.org/10.21462/ijefl.v5i2.302>.
- [31] A. Argianti and H. Retnawati, "Characteristics of Math national-standardized school exam test items in junior high school: What must be considered?," *J. Penelit. dan Eval. Pendidik.*, vol. 24, no. 2, pp. 156–165, 2020, <https://doi.org/10.21831/pep.v24i2.32547>.
- [32] H. H. Dewi, S. M. Damio, and S. Sukarno, "Item analysis of reading comprehension questions for English proficiency test using Rasch model," *Res. Eval. Educ.*, vol. 9, no. 1, pp. 24–36, Jun. 2023, <https://doi.org/10.21831/reid.v9i1.53514>.
- [33] W. Hartono, S. Hadi, R. Rosnawati, and H. Retnawati, "Uji Kecocokan Model Parameter Logistik Soal Diagnosa Kemampuan Matematika Dasar," *JNPM (Jurnal Nas. Pendidik. Mat.)*, vol. 6, no. 1, p. 125, 2022, <https://doi.org/10.33603/jnpm.v6i1.5899>.
- [34] Z. Erdil-Moody and A. S. Thompson, "Exploring motivational strategies in higher education: Student and instructor perceptions," *Eurasian J. Appl. Linguist.*, vol. 6, no. 3, pp. 387–413, 2020, <https://doi.org/10.32601/ejal.834670>.
- [35] S. Handayani, "Analisis uas biologi kelas x dengan teori tes klasik dan item response theory (rasch model)," *Bio-Pedagogi: Jurnal Pembelajaran Biologi*, vol. 11, no. 2, pp. 76-90, 2022, <https://doi.org/10.20961/bio-pedagogi.v11i2.63177>.
- [36] Kartianom, H. Retnawati, K. Hidayati, and A. H. Dewantara, "Understanding Teachers' Experiences with Large-Scale Diagnostic Assessments in Indonesian Islamic Primary Schools: A Phenomenological Study," *Journal of Teaching and Learning*, vol. 20, no. 2, pp. 376-389, 2026, <https://doi.org/10.22329/jtl.v20i1.9250>.

- [37] E. Istiyono, W. S. B. Dwandaru, R. Setiawan, and I. Megawati, "Developing of computerized adaptive testing to measure physics higher order thinking skills of senior high school students and its feasibility of use," *Eur. J. Educ. Res.*, vol. 9, no. 1, pp. 91–101, 2020, <https://doi.org/10.12973/eu-jer.9.1.91>.