Eurasian Journal of Educational Research 106 (2023) 338-350

Eurasian Journal of Educational Research
www.ejer.com.tr

English Final Examination Items Befitting the Criteria: An Item Response Theory Approach

Imtikhanah Rahmah[1]*, Istiyono Edi[2], Widihastuti[3]

ARTICLE INFO

Article History:
Received: 17 May 2023
Received in revised form: 16 December 2023
Accepted: 11 January 2024
DOI: 0.14689/ejer.2023.106.020

Keywords
Final Exam Items, Item Response Theory, Logistic Parameters, Indonesia

ABSTRACT

Purpose: This study aims to describe the characteristics of English Final Examination items at SMA Negeri 5 Malang using the item response theory approach. **Methodology:** This research is quantitative descriptive research with a sample size of 344 students of class XII SMA Negeri 5 Malang. The data was taken from the results of the English final test in the form of multiple-choice questions with 5 choices, totaling 40 items. **Findings:** The results showed that the English final examination test instrument (1) proved to be valid, as indicated by 40 items having a loading factor> 0.3; (2) proved to be reliable as indicated by the reliability coefficient> 0.794; (3) the analysis of the level of difficulty shows that 32 items (80%) have a good level of difficulty category so that they can be used in the next assessment, while the 8 items (20%) do not meet the criteria for a good level of difficulty so they need to be revised or eliminated; and (4) the analysis of differentiating power shows an average index of 0.518 and the index of differentiating power of 40 items (100%) is in the range of 0 to +2. This research is limited to item analysis with a limited sample. **Implications for Research and Practice:** The implications of the study lie in the fact that a wider sample can be taken to be able to analyze up to 3PL. Moreover, the items are only focused on multiple choice, they have not explored the analysis of the quality of the descriptive items which are actually a widely chosen option in student language assessment. Future research is expected to be able to analyze English test items in the form of a descriptive test on a wide scale.

[1] Yogyakarta State University, Indonesia
[2] Yogyakarta State University, Indonesia
[3] Yogyakarta State University, Indonesia
* Corresponding author Email: rahmahdwi.2023@student.uny.ac.id

Introduction

Assessment of learning is very important because it involves assessing the learning process and outcomes (Rios & Guo, 2020; Supena, Darmuki, & Hariyadi, 2021) asserts that evaluation is a very important component in order to assess the level of progression of learning that has been achieved. In other words, evaluation helps to assess the extent to which the learning objectives have been achieved. It helps both teachers and students understand whether learning has taken place effectively. The assessment of learning is supposed to provide a comprehensive description of the various things related to students and teachers who become the subject of education, as well as the stakeholders at the level of praxis of education (Mekonen & Fitiavana, 2021). Tests are constructed to evaluate the extent to which learners have achieved learning objectives. Most tests are based on the instructional material itself, however, a few tests aim to diagnose the performance levels of the learners, to evaluate the learners' capacity to accomplish both the learning and performance objectives (Martha et al., 2021).

This dichotomy can be studied with the Item Response Theory approach (Muranaka, Fujino, & Imura, 2023; Wilson, 2023). The Item Response Theory (IRT) is a blend of educational measurements and psychometrics (Swaminathan, Hambleton, & Rogers, 2006). It makes a comprehensive use of statistical methods such as test development, project analysis, and computer-adaptive testing (Van Der Linden & Hambleton, 1997). All these methods employ mathematical models to examine the relationship between learners' inherent properties and their manifestations (outcomes, responses or performance in a test). When a learner takes the test, the test items and learners' inherent properties are linked together (Baker & Kim, 2004). IRT encompasses all other elements like perception and attitudes and measure them on a continuum. The IRT approach also determines items to be used in a psychometric test. Each item helps measure some aspect of students' ability (Bean, 2022; Embretson & Reise, 2013). The IRT model can be effectively applicable in all types of assessment e.g., psychological, educational and health tests. It can also be used to design and improve scales or measures by including highly discriminative items that help measure accuracy and reduce the burden of answering long questionnaire

There are often three parameters captured in the IRT approach (DeMars, 2010). The parameter model that includes the level of difficulty (bi), known as 1st parameter logistic (1PL); the parameter model that includes both the level of difficulty (bi) and the discriminating power (ai) is known as 2PL; and the parameter model that includes the level of difficulty (bi), the discriminating power (ai), and the pseudo guess (ci) is known as 3PL. Additionally, the analysis of items is also carried out by proving validity, estimating reliability, and calibrating with the item response theory approach. So how is a test instrument said to be valid? A priori, it can be stated that a valid test instrument is an instrument that can be used to measure what should be measured. Reynolds, Altmann, and Allen (2021) stated that a good-criteria instrument is an instrument that satisfies the requirements of validity and reliability. Meanwhile, Ramadhan et al. (2019) believed that the instrument is considered to be valid if it can be proven that the instrument accurately measures the ability of students in accordance with the competencies measured and that the test instrument must be able to prove its validity in measuring student abilities.

In addition to validity, the instrument must also satisfy reliability criteria. Reliability is the degree of consistency between two measured scores on the same object, despite using

different measuring instruments and different scales (Reynolds et al., 2021), reliability can be used to determine the consistency of the measuring instrument, whether the measuring instrument remains consistent if it produces the same results despite repeated measurements. The reliability assumption is fulfilled if the observed score has a high relationship with the actual score (Allen & Wilson, 2006). It is also stated by Retnawati (2017) that a quality instrument will always produce a higher value of information than measurement error. In a test instrument, satisfying the assumptions of validity and reliability cannot be without the criteria of difficulty, differentiation, and distractors (Anita, Wu, & Abdillah, 2023). This can be analyzed with the Rasch model, which is a modern assessment theory that can classify item and person calculations in a distribution map (Rozeha, Azami, & Mohd Saidfudin, 2007). This model is also part of item response theory (Thissen & Twaalfhoven, 2001).

The Indonesian education system employs the Final Examination as the highest level of test at the high school level. Teachers, as the people who are most familiar with student development, have more of a responsibility as question writers at the school level. Although in some areas, MGMP (teachers' association) as an organization of teachers in each province makes a blue print or grid of questions, in order to have similar quality questions prepared between one school and another, though it is not enough to ensure quality of tests and questions sets in each school. The English final exam, as one of the most important tests at the high school level, must satisfy the eligibility criteria (Thissen & Thissen-Roe, 2020).

Until now, the examination instrument has only been examined by peer-checking. Therefore, considering the important position of the final exam, this research was conducted with the aim of analyzing and describing the characteristics like validity, estimating reliability, and calibrating the practicality and usability of the items in the English language subject used in the school exam at SMA Negeri 5 Malang.

Literature Review

- *Characteristics of Good Test Instruments*

Test instruments are one of the methods to assess the level of a person's ability indirectly, through a person's response to a stimulus or question (Mardapi, 2020). In other words, a test is a way or method to see a student's capability in responding to assigned tasks by using his/her acquisition of a skill or knowledge. There are a few characteristics that determine the quality of test instruments, namely validity, reliability and usability of the items (Nurcahyo et al., 2019).

Validity is first and foremost characteristics of any good test. It refers to the extent to which the test serves its purpose. In other words, validity is the measure of the efficiency at which learners' performance can be measured or the test items can be measured. To judge the validity of any test, it is also required to know the purpose of the test. Having fulfilled the real purpose, the test score can show consistency. The validity of a test can be classified as content validity, concurrent validity, predictive validity and construct validity. Each of these types measures various factors and cater of specific objectives.

Content validity, for instance, refers to the extent to which the test content represents the content of the course, fulfills its objectives and its subject matter. This type is judged in three main domains viz., cognitive, affective and psychomotor. Concurrent validity type refers to the degree to which the test correlates to the criterion of the test and its measurements are acceptable. It may use a statistical tool to correlate and interpret test results. Predictive validity is the type that measures the learners' actual performance in a test and predicts true results. This prediction determines the future outcomes of the learners and validation of their score. Lastly, construct validity focuses on the theoretical traits in the test items such as intelligence, reading comprehension, critical thinking, or mathematical aptitude (Nurcahyo et al., 2019).

Reliability, on the other hand, is required to determine accuracy and consistency, to determine the extent to which a test is consistent, stable and dependent. This characteristic is also the evidence that a test can be taken a number of times, and would give the same result every time. It means that when a student scores 80 marks in a test on a certain day, and on another day, if he takes the same test and scores 30 or 40 marks, the test cannot be called reliable. The inconsistency of test result can also adversely affect learners' scores. Finally, the usability and practicability of a test refers to the extent to which it can be used without much administrative and practical difficulties. This feature therefore requires that a test should be administered with clarity, ease, and uniformity. For this purpose, it requires that the test should be simple, concise, and clear; it should have a time limit, sample questions, and oral instructions (Azwar, 2015).

- *Test Types and Their Planning*

There are several types of tests such as criterion-referenced test, performance test, and attitude survey test. Criterion-referenced tests are most commonly used, as they are used to test all the three learning domains of cognitive, affective, and psychomotor (Krathwohl, 2012). When it evaluates the cognitive domain, for instance, it assesses the recall or recognition of facts, procedural patterns, and concepts that serve in the development of intellectual abilities and skills. The testing of these abilities and skills are often measured with a written test or a performance test. A performance test evaluates the psychomotor domain that involves physical movement, coordination, and use of the motor-skill areas. It measures speed, precision, distance, procedures, or techniques in execution. Lastly, attitude survey test evaluates the affective domain that addresses the manner in which we deal with things emotionally, such as feelings, values, appreciation, enthusiasms, motivations, and attitudes. Attitudes are not observable; therefore, a representative behavior must be measured.

Unlike two other types, a performance test helps the learner to demonstrate a skill that has been instructed in a training program. Performance tests are criterion-referenced when they require the learner to demonstrate the required behavior stated in the objective. There are three critical factors to draft a performance test: first, the learner must know what behaviors (actions) are required in order to pass the test, for which s/he must take adequate practice and coaching; second, all test equipment must in good working condition prior to the test, for which prior planning and necessary resources availability are required; and third, the test administrator must know what behaviors are to be looked for and how they are rated. The performance evaluation is best measured when each step of the task and all parameters are completed in the given time.

Whatever type of the test, it must be well planned in advance. If not planned, a few test items may be over or under represented, while others may not even be touched. Sometimes, the test items are built on topics that are easy ones and does not require any challenge or hard work as preparation. A well-planned test has open-ended questions, which may have an unlimited answer; closed- ended questions with alternate responses such as yes/no or true/false or multiple-choice questions. A few tests are descriptive which requires long answers in the form of paragraphs or essays. A major challenge in open or closed-ended questions is that they emphasize isolated bits of information and measures only the learner's ability to recognize the right answer, but not the ability to recall or reproduce the right answer. Likewise, the major concern in composition tests requiring paragraphs and essays is the wide variance in their grades which instructors follow. Whatever it the level of difficulty in these types of tests, learners answer the tests according to the level of their ability. If there are many students who answer a question correctly, then the question tends to be easy, and vice versa. Questions that are considered too easy or too difficult are not necessarily optimal in identifying the high and low abilities of students. If the discriminative capacity of a question item is weak, then the information provided by the item is inaccurate. Meanwhile, the strength of distraction will affect the function of the multiple choices.

- *Item Response Theory*

The item response theory (IRT), also known as the latent response theory, comprises such mathematical models that explain the relationship between latent traits (unobservable characteristic or attribute) and their manifestations (i.e. observed outcomes, responses or performance) (DeMars, 2010; Swaminathan et al., 2006). Such models help to establish a link between the properties of items on a test instrument and the learners responding to these items, as well as the specific trait being measured. IRT also ensures that the latent traits such as stress, knowledge, attitudes and items of a measure are organized in an unobservable continuum. Therefore, its main purpose focuses on establishing the individual's position on that continuum (Embretson & Reise, 2013; Van Der Linden & Hambleton, 1997).

The IRT not only helps in the test development and determining the scoring methods despite the difficulty levels of the tests, it also maximizes people's ability to distinguish between latent traits (Baker & Kim, 2004). The unit of analysis of IRT models is the item, which can be utilized to compare different measured items, as long as they measure the same underlying structure. Furthermore, they can be used in different functions of items to assess why calibrated and tested items always behave differently between groups. This can lead to studies identifying agents responsible for differences in responses and linking them to population characteristics. Last, but not the least, the IRT demonstrates the probability that a person with certain sets of knowledge and skills will possess a potential to perform at a particular level.

Methodology

- *Research Design*

This study adopted a quantitative and descriptive research design, which aimed to describe the situation that occurred in accordance with the object under study. The research

was conducted by using secondary data in the form of score obtained by XII grade students of SMA Negeri 5 Malang, Indonesia who took the final Examination in English subject.

- *Sampling*

The purposive sampling technique was adopted to identify the 344 XII grade students of SMA Negeri 5 Malang who appeared in the final exam. The data was configured according to the level of difficulty of questions into easy, moderate and difficult with the highest percentage in moderate questions as much as 50%, easy 30%, and difficult 20%.

- *Research Instrument and Procedure*

The test instrument comprised a multi-item questionnaire built on the item response theory principles. It contained multiple choice, matching, true false and complex multiple-choice questions totaling 40 questions, with each question having a value of one score.

- *Data Analysis*

The dichotomous data was analyzed to examine its construct validity through exploratory factor analysis (EFA) and the reliability of the items was estimated using Cronbach's alpha formula with the support of SPSS V.25 software. Meanwhile, for item calibration, the JMetrik software was used with the item response theory approach. Item characteristics were determined according to the criteria of the 2 PL parameter index, which included difficulty level (bi) and differentiation power (b2). Thus, the English questions tested in the implementation of the final examination were considered to fulfill the quality criteria if the index is in the range of -2 to +2, the discriminating parameter (ai) in the range of 0 to +2, and the pseudo guessing parameter (ci) in the range of 0 to 1/k (DeMars, 2010).

Results and Discussion

- *Validity and Reliability of the English Final Test Instrument*

In proving the construct validity using exploratory factor analysis (EFA), there are certain assumptions that are required. First, through the chi-square value in the Bartlett test and the Kaiser Meyer Olkin measure of sampling adequacy (KMO-MSA), we can determine the assumption of sample adequacy. While the eigenvalue and scree plot represent the number of dominant factors generated in the instrument, and the component matrix contains loading factors. In the analysis of the construct validity of the English final test instrument, sample adequacy is shown by the KMO-MSA and Bartlett test in Table 1.

Table 1

KMO and Bartlett's Test.

Kaiser-Meyer- Olkin Measure of Sampling Adequacy		**0.823**
Bartlett's test of Sphericity	Approx Chi-square	2356.203
	df	780
	Sig	.000

Table 1 shows that the English final test instrument data tested on 344 students at SMA Negeri 5 Malang obtained a KMO value of 0.823. The KMO value, which is greater than 0.50, indicates that the sample size has been met for factor analysis (Dani, Singhai, & Anand, 2023). Meanwhile, the p value of 0.000 on Bartlett's Test of Sphericity shows the correlation between items. The value listed is <0.05 accordingly the items are correlated. So, it can be concluded that the data is sufficient for factor analysis and proof of construct validity.

The loading factor in the component matrix shows that all items have a loading factor> 0.3. So that the English final test instrument can be considered valid, which measures the dimensions to be measured. The lowest loading factor is on item number 38, which is 0.303. While the highest loading factor is at number 12, which is 0.794. There are 22 items that measure on the 1st factor, while the other 18 items are spread across other factors. This shows that there is 1 dominant factor measured by the English final test instrument, which is the 3rd factor.

In addition, based on the reliability estimation by using Cronbach's Alpha, a reliability coefficient of 0.794 was calculated as seen in Table 2. These results show that the instrument was categorized as reliable because it had a reliability coefficient > 0.7. The instrument satisfies the reliability criteria if it obtains a reliability coefficient> 0.7 (Azwar, 2015).

Table 2

Alpha Cronbach Reliability Estimation,

Reliability Statistics	
Cronbach's alpha	No of items
.794	40

- *Assumption of Unidimensional and Local Independence*

The item calibration in this research uses the item response theory approach, which requires an assumption test, including assumptions of uni-dimensionality, local independence, and parameter invariance. The assumption of uni-dimensionality can be known in two ways, i.e. by looking at the eigenvalue or the steepness of the scree plot (Retnawati, 2017). The eigenvalue table and scree plot of the English final test instrument can be found in Table 3.

There are 15 components with eigenvalues > 1, which are shown in Table 3. It indicates that the 15 components have been able to explain as much as 59.7% of the variation of the entire instrument. In addition, it can also be seen that the eigenvalue of the first component is 15.656, which is quite far from the other factors' eigenvalues. It shows that the English final test instrument measures 1 dominant factor. If the first eigenvalue has a multiple value of the second component and the eigenvalue between the next components is almost the same, then the assumption of uni-dimensionality is fulfilled (Susetyo, 2015).

Table 3

Eigenvalue of English Final Test Instrument.

Total Variance Explained						
	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Factor	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	6.262	15.656	15.656	2.488	6.220	6.220
2	1.566	3.914	19.570	1.273	3.183	9.403
3	1.528	3.820	23.390	3.919	9.799	19.201
4	1.452	3.630	27.020	.914	2.285	21.486
5	1.409	3.522	30.543	.814	2.036	23.522
6	1.322	3.306	33.848	.848	2.120	25.642
7	1.291	3.226	37.075	.705	1.763	27.405
8	1.272	3.181	40.256	.643	1.608	29.013
9	1.226	3.064	43.320	.705	1.762	30.775
10	1.180	2.951	46.271	.515	1.289	32.063
11	1.158	2.895	49.165	.554	1.385	33.448
12	1.109	2.772	51.938	.517	1.293	34.741
13	1.060	2.651	54.589	.452	1.131	35.872
14	1.046	2.616	57.205	.407	1.017	36.889
15	1.001	2.502	59.707	.400	1.000	37.889

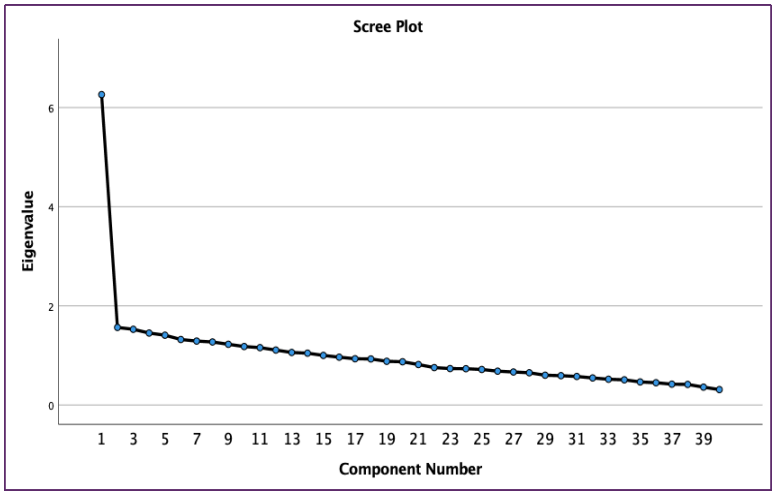

Figure 1. *Scree Plot of English Final Test Instrument.*

Figure 1 presents the scree plot results to reinforce the conclusions from the eigenvalues shown in Table 3. It appears that the first component and the second component make a long steepness, while the second component to the third component is short and sloping. The steepness of the scree plot indicates the number of dominant dimensions and the ramps do not indicate the existence of dimensions (Retnawati, 2017). Thus, based on the eigenvalue and reinforced by the scree plot formed, it can be concluded that the English final test instrument is unidimensional.

Furthermore, the assumption of local independence must also be satisfied in the item response theory approach. Local independence is an assumption that requires the score of a test item to be independent of other items (Falani, Akbar, & Naga, 2020). According to DeMars (2010), the assumption of local independence will be fulfilled if a test is proven to be unidimensional (2018). Thus, the English final test instrument satisfies the assumption of local independence because it has been proven to be unidimensional.

Descriptive Statistics

Swaminathan et al. (2006) revealed that the item response theory approach can be used if it is a fit between the logistic model and the test data. The item response theory approach offers three logistic models, i.e. 1 PL contains the difficulty parameter (bi), 2 PL contains the difficulty parameter (bi) and discriminating power (ai), and 3 PL contains the difficulty parameter (bi), discriminating power (ai), and pseudo guess (ci). As more parameters are used, so the more detailed the logistic model informs the test taker's ability. In this study, since there are only 344 data to be analyzed, the approach used is only the 2PL logistic model.

The compatibility of the items against the logistic model can be determined by comparing the chi square ($\chi 2$) calculation with the chi square ($\chi 2$) table (Retnawati, 2017). In addition, item suitability can also be viewed from the probability value (P value), if the P value> alpha (0.05) then the item is suitable for the model. From the data analyzed, the results of the English final test instrument suitability test through p-value> 0.05 indicated that 11 items had p-value < 0.05 and 29 items had p-value < 0.05.

- *Assumption of Parameter Invariance*

Furthermore, an analysis is carried out to obtain parameter invariance. This is an assumption that requires the parameters of an item to be independent of test takers and vice versa (Duskri, Kumaidi, & Suryanto, 2014). It is clarified by Jumini and Retnawati (2022) that the parameter invariance can be determined based on the invariance of item parameters and test participants' abilities. Providing evidence of parameter invariance can also be carried out by cross-correlating between groups, if the groups have a strong relationship, it can be concluded that the parameters are invariant (Susetyo, 2015). Proving the assumption of parameter invariance in this research is viewed from the correlation of the residual data, where the results obtained are not exceeding -0.2 to 0.2. And the results of the analysis, the lowest correlation value is -0.9158, and the largest is 1.

Item Characteristics

The item calibration of the English final test instrument of SMA Negeri 5 Malang utilizing the item response theory approach with the 2PL model produces the parameters of difficulty (bi) and distinguishing power (ai). The calibration shows that the item with the lowest difficulty index is item number 21 with an index of -2.25, while the highest or most difficult item is item number 37 with an index of 3.31. According to Susetyo (2015) that items with a range close to logit +2 then the item tends to be more difficult, otherwise if the item is close to -2 it is relatively easy, and in the range $-1.0 < b < +1.0$ the item is classified as moderate. Meanwhile, items that have a difficulty index outside the range can be corrected or eliminated (Jumini & Retnawati, 2022).

Among the 40 questions tested, there are 20% or about 8 items that must be revised because they have an unfavorable difficulty index, which is out of the index range of -2 to +2, i.e. on numbers 1, 10, 11, 16, 21, 27, 33, 37. While the other 80% have been categorized as good (32 items). When analyzed in more detail, there are 5 items or as many as 12.5% classified as difficult, 3 items (7.5%) classified as easy, and the remaining 32 items classified as medium. So, it can be concluded that the overall level of difficulty of the final English items is classified as medium with an average difficulty index of -0.02.

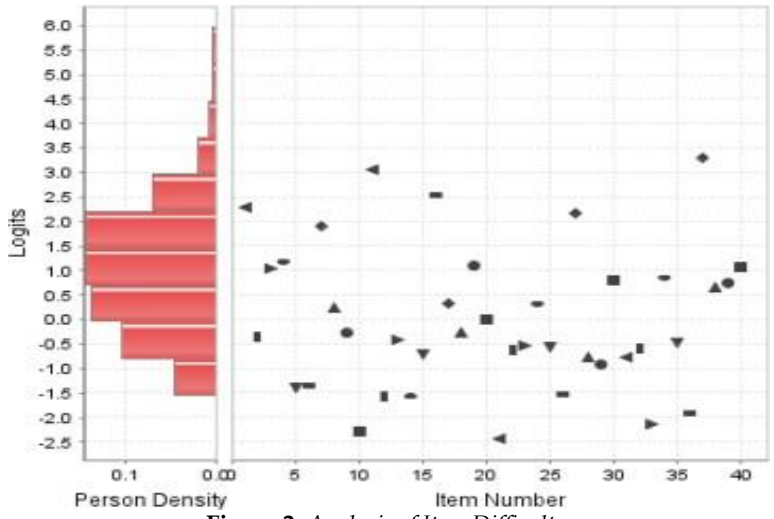

Figure 2. *Analysis of Item Difficulty.*

In addition, the calibration results also indicate the characteristics of item discriminating ability. DeMars (2010) explains that the discriminating ability has a range of 0 to +2. Overall, the characteristics of the discriminating ability of the English final test instrument are classified as good with an average index of 0.518 and the discriminating ability index of 40 items (100%) is in the range of 0 to +2. The lowest discriminating ability was found in item number 11 with an index of 0.15, while the highest index was found in item number 8 with an index of 1.42. So, it can be concluded that the final English test instrument of SMA Negeri 5 Malang is able to discriminate between high-ability and low-ability students.

Conclusion and Recommendations

Based on the research results presented previously, it can be concluded that the final English test instrument of SMA Negeri 5 Malang proved to be valid and reliable. The validity of the test instrument is indicated by the items, all 40 of which have a loading factor> 0.3. While its reliability is proven by the reliability coefficient> 0.7, that is 0.794. In the difficulty analysis, the final English test instrument of SMA Negeri 5 Malang is classified as moderate with an average difficulty index of -0.02, with details of 32 items (80%) having a good level of difficulty so that they can be used in the next assessment, while 8 items (20%), i.e. item numbers 1, 10, 11, 16, 21, 27, 33, and 37 do not fulfill the criteria for a good level of difficulty so they need to be revised or eliminated.

Moreover, the differentiating power analysis shows that the English final test instrument is classified as good with an average index of 0.518 and the discriminating factor index of 40 items (100%) is in the range of 0 to +2. So, it can be concluded that the English final test instrument of SMA Negeri 5 Malang is able to distinguish high ability and low ability students. This research is limited to item analysis with a limited sample. In the future, it is expected that a wider sample can be taken to be able to analyze up to 3PL. Moreover, the items are only focused on multiple choice, they have not explored the analysis of the quality of the descriptive items which are actually a widely chosen option in student language assessment. Future research is expected to be able to analyze English test items in the form of a descriptive test on a wide scale.

References

Allen, D. D., & Wilson, M. (2006). Introducing multidimensional item response modeling in health behavior and health education research. *Health Education Research, 21*(suppl_1), 73-84. https://doi.org/10.1093/her/cyl086

Anita, R., Wu, W., & Abdillah, M. R. (2023). Developing a Scale of Ethical Responsibility (SER): A Multi-Dimensional Instrument for Fintech Professionals. *Social Indicators Research, 170*, 1007-1033. https://doi.org/10.1007/s11205-023-03231-5

Baker, F. B., & Kim, S.-H. (2004). Item Response Theory: Parameter Estimation Techniques. *Journal of Educational Measurement, 31*, 3. http://www.jstor.org/stable/1435270

Bean, G. J. (2022). Assessing Differential Item Functioning and Differential Test Functioning in an Academic Motivation Scale using Item Response Theory methods. *International Journal of School Social Work, 8*(1), 2. https://doi.org/10.4148/2161-4148.1096

Dani, S., Singhai, M., & Anand, M. M. (2023). Blended mode of Learning: New Normal for 21st Century Learners. *The Online Journal of Distance Education and e-Learning, 11*(4), 3030-3041. https://tojdel.net/journals/tojdel/articles/v11i04/v11i04-04.pdf

DeMars, C. (2010). *Item Response Theory*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195377033.001.0001

Duskri, M., Kumaidi, K., & Suryanto, S. (2014). Pengembangan tes diagnostik kesulitan Belajar matematika di SD. *Jurnal Penelitian dan Evaluasi Pendidikan, 18*(1), 44-56. https://doi.org/10.21831/pep.v18i1.2123

Embretson, S. E., & Reise, S. P. (2013). *Item Response Theory for Psychologists* (1st ed.). Psychology Press. https://doi.org/10.4324/9781410605269

Falani, I., Akbar, M., & Naga, D. S. (2020). The Precision of Students' Ability Estimation on Combinations of Item Response Theory Models. *International Journal of Instruction, 13*(4), 545-558. https://doi.org/10.29333/iji.2020.13434a

Jumini, J., & Retnawati, H. (2022). Estimating Item Parameters and Student Abilities: An IRT 2PL Analysis of Mathematics Examination. *AL-ISHLAH: Jurnal Pendidikan, 14*(1), 385-398. https://doi.org/10.35445/alishlah.v14i1.926

Krathwohl, D. R. (2012). Stating objectives appropriately for program, for curriculum and for instructional materials development. In *Educational Objectives and the Teaching of Educational Psychology* (pp. 188-206). Routledge. https://doi.org/10.4324/9780203807408-11

Mardapi, D. (2020). Assessing Students' Higher Order Thinking Skills Using Multidimensional Item Response Theory. *Problems of Education in the 21st Century, 78*(2), 196-214. https://doi.org/10.33225/pec/20.78.196

Martha, A. S. D., Junus, K., Santoso, H. B., & Suhartanto, H. (2021). Assessing undergraduate students'e-learning competencies: A case study of higher education context in Indonesia. *Education Sciences, 11*(4), 189. https://doi.org/10.3390/educsci11040189

Mekonen, Y. K., & Fitiavana, R. A. (2021). Assessment of learning outcomes in higher education: Review of literature. *Assessment of Learning Outcomes in Higher Education: Review of literature, 71*(1), 69-76. https://doi.org/10.47119/IJRP100711220211766

Muranaka, S., Fujino, H., & Imura, O. (2023). Evaluating the psychometric properties of the fatigue severity scale using item response theory. *BMC Psychology, 11*(1), 1-11. https://doi.org/10.1186/s40359-023-01198-z

Nurcahyo, F. A., Azwar, S., Martani, W., & Kartowagiran, B. (2019). Development and Psychometric Properties of Pictorial Vocational Interest Inventory for Indonesian Adolescents. *Electronic Journal of Research in Educational Psychology, 17*(47), 213-236. https://doi.org/10.25115/ejrep.v17i47.2132

Ramadhan, S., Mardapi, D., Prasetyo, Z. K., & Utomo, H. B. (2019). The development of an instrument to measure the higher order thinking skill in physics. *European Journal of Educational Research, 8*(3), 743-751. https://doi.org/10.12973/eu-jer.8.3.743

Retnawati, H. (2017). Learning trajectory of item response theory course using multiple software. *Olympiads in Informatics, 11*, 123-142. https://doi.org/10.15388/ioi.2017.10

Reynolds, C. R., Altmann, R. A., & Allen, D. N. (2021). Item Analysis: Methods for Fitting the Right Items to the Right Test. In *Mastering Modern Psychological Testing: Theory and Methods* (pp. 263-289). Springer. https://doi.org/10.1007/978-3-030-59455-8_7

Rios, J. A., & Guo, H. (2020). Can culture be a salient predictor of test-taking engagement? An analysis of differential noneffortful responding on an international college-level assessment of critical thinking. *Applied Measurement in Education, 33*(4), 263-279. https://doi.org/10.1080/08957347.2020.1789141

Rozeha, A. R., Azami, Z., & Mohd Saidfudin, M. (2007). Application of Rasch Measurement in Evaluation of Learning Outcomes: A case study in electrical engineering. In *Regional Conference on Engineering Mathematics, Mechanics, Manufacturing & Architecture (EM3ARC)*. https://www.semanticscholar.org/paper/a264b058246d68b10c1de1b1c52f4252170604c2

Supena, I., Darmuki, A., & Hariyadi, A. (2021). The Influence of 4C (Constructive, Critical, Creativity, Collaborative) Learning Model on Students' Learning Outcomes. *International Journal of Instruction, 14*(3), 873-892. https://doi.org/10.29333/iji.2021.14351a

Susetyo, H. (2015). Contestation between State and Nonstate Actors in Zakah Management in Indonesia. *Jurnal Syariah, 23*(3), 517-546. https://doi.org/10.22452/js.vol23no3.7

Swaminathan, H., Hambleton, R. K., & Rogers, H. J. (2006). Assessing the Fit of Item Response Theory Models. *Handbook of statistics, 26*, 683-718. https://doi.org/10.1016/S0169-7161(06)26021-8

Thissen, D., & Thissen-Roe, A. (2020). Factor score estimation from the perspective of item response theory. In *Quantitative Psychology: 84th Annual Meeting of the Psychometric Society, Santiago, Chile, 2019 84* (pp. 171-184). Springer. https://doi.org/10.1007/978-3-030-43469-4_14

Thissen, W. A. H., & Twaalfhoven, P. G. J. (2001). Towards a conceptual structure for evaluating policy analytic activities. *European Journal of Operational Research, 129*(3), 627-649. https://doi.org/10.1016/S0377-2217(99)00470-1

Van Der Linden, W. J., & Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. In *Handbook of modern item response theory* (pp. 1-28). Springer. https://doi.org/10.1007/978-1-4757-2691-6_1

Wilson, M. (2023). A Review of: "Constructing Measures: An Item Response Modeling Approach". *International Journal of Testing, 8*(2), 197-201. https://doi.org/10.1080/15305050802007117