ENGLISH EDUCATION
English Journal for Teaching and Learning
Vol. 12 No. 02 Desember 2024 pages 182 - 196
https://jurnal.uinsyahada.ac.id/index.php/

Personal Fabricated-English Items' Quality: Classical Test and Item Response Theories

Try Mahendra Siregar*[1], Riandry Fadilah Nasution[2], Putri Nurul A'la[3]

[1,3] Syiah Kuala University, Indonesia, [2] The University of Manchester, United Kingdom
e-mail: *[1]trymahendrasiregar@usk.ac.id, [2]riandry.nasution@postgrad.manchester.ac.uk, [3]putrinurul.ala@usk.ac.id

Abstract Item Analysis is used to determine the quality of test items, whether applicable or not applicable for the test takers' ability assessment. Owing to that, our research attempts to measure the quality of personal fabricated English items for 8th grade students under the Classical Test Theory (CTT) and Item Response Theory (IRT) by Rasch models with Quest.exe application. We adopted reliability, item difficulty, discrimination power, and distractor effectivity. 30 items with multiple-choice format were handed out to 46 students and result showed that the items are reliable with 0.69 CTT and 1.0 IRT values, and the item difficulties are also varied: 12, 14, and 4 based on CTT categorizations and index easy, moderate, and difficult, while IRT demonstrated similar results. There is only 1 item inadequate to differentiate students' ability, revision required; furthermore, 17 out of 30 items have effective distractors. This research is expected to contribute to Item analysis and Quest.exe demonstration for the same purposes.
Keywords: *Classical Test Theory; Item Response Theory; Quest.exe, EFL*

Abstrak Analisis Soal digunakan untuk mengetahui kualitas soal tes, apakah dapat diterapkan atau tidak untuk penilaian kemampuan peserta tes. Oleh karena itu, penelitian kami mencoba mengukur kualitas soal bahasa Inggris buatan pribadi untuk siswa kelas 8 berdasarkan *Classical Test Theory* (CTT) dan *Item Response Theory* (IRT) dengan model Rasch melalui aplikasi Quest.eexe. Kami mengadopsi keandalan, kesukaran item, kekuatan diskriminasi, dan efektivitas pengalih perhatian, mengikuti kedua teori tersebut. 30 soal dengan format pilihan ganda dibagikan kepada 46 siswa dan hasil penelitian menunjukkan bahwa soal-soal tersebut reliabel dengan nilai CTT 0,69 dan IRT 1,0, serta tingkat kesulitan soal juga bervariasi: 12, 14, dan 4 berdasarkan kategorisasi CTT dan indeks mudah, sedang, dan sulit, sedangkan IRT menunjukkan hasil yang serupa. Hanya terdapat 1 butir soal yang kurang mampu membedakan kemampuan siswa dan butir soal tersebut perlu direvisi; lebih jauh lagi, 17 dari 30 item mempunyai pengecoh yang efektif. Penelitian ini diharapkan dapat berkontribusi pada analisis Item dan demonstrasi Quest.exe pada tujuan yang sama.
Kata Kunci: *Teori Tes Klasik; Teori Butir Respon; Quest.exe*, EFL

INTRODUCTION

The legislation No. 20, 2003, article 39 paragraph 2 concerning the National Education System states that educators are professional staff in charge of planning and implementing the learning process, assessing learning outcomes, conducting guidance and training, as well as conducting research and service to the community especially to educators in higher education level (Law of the Republic of Indonesia Number 20, 2003). One of the competencies, that teachers should have, is the ability to evaluate students' learning processes or learning outcomes. The evaluation is sometimes mistaken for the test. However, while these terms are distinct, they are inextricably linked.

An evaluation is defined as a process or act of determining the significance of an item (Suarga, 2019). Asrul et al. (2014) states that learning evaluation is basically not just assessing learning outcomes, but also the process that educators and students have done during learning activities and process. Compared to evaluation, a test, by Bachman (1990, as cited in Mahmoodi-shahrebabaki, 2018) is a measurement instrument designed to obtain specific sample of individuals behaviour or to qualify certain characteristics based on explicit procedures. The test is an assignment or a set of tasks in the form of items or questions that learners must complete. The results are utilized to make specific inferences about the learners own learning results as well as to diagnose the compatibility of prepared and taught materials or the test items altogether.

As mentioned earlier, the result of test is used to determine the quality of the test; whether the quality of given test is fit or misfit for the test takers or learners. This activity is called items analysis. In test, several characteristics of the items are known. There are three characteristics of test namely; difficulty level, discrimination power and distribution of answers or functional of items distractors. Three of these characteristics are in set to determine the quality of the items, so if any of these does not meet the requirements, the quality of the items will decrease. Zuriyati (2016, as cited in Fitrianawati, 2017) states the aims of item analysis are 1) to determine compatible or not the items for learners, 2) to improve the quality of items through the three characteristics; difficulty level, discrimination and distractor appropriateness, 3) to increase the validity and reliability of the items, 4) and as a fundamental to revise irrelevant materials being taught through how many correct responses by learners.

In order to qualify an item, there are two approaches are commonly applied for the analysis; the first, Classical Test Theory (CTT) upbringing two main statistics on item

facility index (the proportion of correctly answered items) and discrimination index (the Pt-Biserial of students' performance and total test score), the second, Item Response Theory (IRT) that concerning both item statistics and students' ability on single item to entire test performance (Erfan et al., 2020; Goolamally, 2019; Heppi Yuslita, 2016; Susdelina, 2018). With those approaches, some research regarding item assessments have been done by deploying the both or one of those theories such as concerning validity and reliability of Item with IRT (Huang, et.al, 2023; Siri & Freddano, 2011), qualifying item level of difficulty, discrimination power, and/or distractor function with CTT (Danuwijaya, 2018; Hartati & Yogi, 2019; Karim et al., 2021; Khairuddinalfath, 2019; Ningsih & Widowati, 2021; Suek, 2021), and analyzing item level of difficulty, discrimination power, and/or distractor function with CTT and IRT (Ashraf & Author, 2020).

From all aforementioned related research, the CTT appeared to be more popular among the scholars than the IRT, therefore, to maximize a small gap, this research will apply both theories (CTT and IRT specified on Rasch Model) for the analysis to degree how applicable the fabricated items for students for a better revision. Additionally, the Quest.exe application will be utilized to attain the statistical result and validation. The findings of this research are thought to be theoretically and practically essential to item analysis for teacher individual analysis under the CTT and IRT Rasch Models, as well as demonstrating to teachers how Quest.exe functions as a tool for analysis.

METHOD

The present research was a descriptive quantitative research project that used Classic Test Theory (CTT) (Goolamally, 2019; Heppi Yuslita, 2016; Susdelina, 2018) and computer applications with the Item Response Theory (IRT) with Rasch model (Erfan et al., 2020; Goolamally, 2019; Heppi Yuslita, 2016; Susdelina, 2018) to explain information about the overall quality of the test items. According to Creswell (2002; 2012), quantitative research is the process of collecting, analyzing, interpreting, and writing the results of research. The current research data collection involved testing 46 participants, all of whom were in their eighth grade at junior high school. The students had been confirmed to have studied comparative degree, past tense, and report text discussion.

The research was done in phases, as follows: 1) developing a test table specification, namely generating a blueprint relating to the items to be tested. The

blueprint categories included basic competencies, indicator of lessons, regular material (subject discussion), test format (multiple choices with over all 30 items), test technique, and item numbers, 2) Arranging the Items and checking language suitability for items is done by requesting language experts, particularly in English as the target language. Then, the revision of the Language in Items is done, namely, by following the corrector's suggestions and comments, 4) testing items with the participants directly and 5) Evaluating the appropriateness of the items with the CTT and IRT Rasch models is by utilizing Quest.exe. The following tables describe the indexing used to quantify the appropriateness of the items.

Table 1.
Reliability in CTT & IRT
(Goolamally, 2019)

Reliability Categories in CTT	Coefficient	Reliability Categories in IRT	Coefficient
Very High	0.90 – 1.0	Reliable	> 0.70
High	0.70 – 0.89		
Satisfactory	0.40 – 0.69		
Low	0.20 – 0.39	Not-reliable	> 0.70
Very Low	0.0 – 0.19		

Table 2.
Item Level of Difficulty in CTT
(Robert & Hagen, 2009, as cited in Yuslita, 2016; Susdelina, 2018)

Categories in CTT	Difficulty Index	Categories in IRT	Measurement Value
Difficult	< 0.30	Very Difficult	> 1.0
Moderate	0.30 – 0.70	Difficult	0 – 1.0
Easy	> 0.70	Easy	-1 – 0
		Very Easy	< -1

Table 3.
Discrimination Power in CTT
(Susdelina, 2018)

Categories in CTT	Measurement Value
Very High	$0.70 < D \leq 1$
High	$0.40 < D \leq 0.70$
Moderate	$0.20 < D \leq 0.40$
Low	$0 < D \leq 20$
Very Low	$D \leq 0$

Furthermore, as for the discrimination power in the IRT Rasch Model, the value utilized is in the range of 0.77 to 1.33, so when the value is less than the expected range, it indicates the participants are not compatible to answer the questions (Erfan et al., 2020; Goolamally, 2019; Susdelina, 2018). Additionally, the distractor's effectiveness is

measured by the number of correct responses. A good distractor is ideally chosen by 5% of the total participants; therefore, a lower proportion means the distractor needs to be adjusted or revised.

RESULT

Following to this research, there are two approaches to consider whether or not the EFL items reliable for testing. The result of reliability by CTT in Quest.exe output can be seen from Reliability of Estimate in the figure below:

Figure.1
Item Estimates (Thresholds)

```
TPA
-------------------------------------------------------------
Item Estimates (Thresholds)
all on all (N = 46 L = 30 Probability Level= .50)
-------------------------------------------------------------

Summary of item Estimates
=========================

Mean                                  .00
SD                                   1.02
SD (adjusted)                         .95
Reliability of estimate               .88

 Fit Statistics
----------------

 Infit Mean Square          Outfit Mean Square

     Mean     1.00              Mean       .98
     SD        .07              SD         .14

       Infit t                    Outfit t

     Mean       .11              Mean       .06
     SD         .52              SD         .49

   0 items with zero scores
   0 items with perfect scores
-------------------------------------------------------------
```

The reliability of the tested items is at 0.88 estimation, these items are categorized in high reliability. The IRT Rasch model, however, provides two type of FIT statistics (INFIT Mean Square Statistics (MNSQ) is the sensitivity to an expected response on the items and OUTFIT is the outline sensitivity. These both FIT statistics is applied for reliability value of participants and items with approximation on Cronbach – α is .0.70. For the result of this research, the FIT MNSQ is in 1.00 with 0.7 Standard Deviation (SD), so, the items are reliable (higher than 0.70).

Additionally, in figure 2, there is no zero or perfect scores for all items. The reliability value on smaller scale of cases or person has been concluded below:

Figure 2.
Summary of Case Estimates

```
TPA
------------------------------------------------------------------------
Case Estimates
all on all (N = 46 L = 30 Probability Level= .50)
------------------------------------------------------------------------

Summary of case Estimates
-------------------------

Mean                                .50
SD                                  .79
SD (adjusted)                       .66
Reliability of estimate             .69

 Fit Statistics
----------------

 Infit Mean Square          Outfit Mean Square

    Mean      1.00             Mean       .98
    SD         .18             SD         .25

      Infit t                    Outfit t

    Mean       .04             Mean       .01
    SD         .96             SD         .73

   0 cases with zero scores
   0 cases with perfect scores
------------------------------------------------------------------------
```

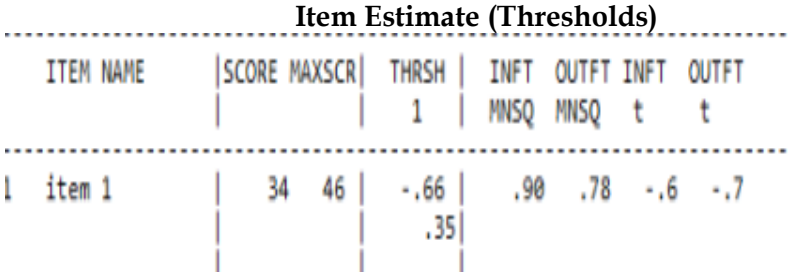
The reliability value on smaller cases in person is in 0.69, it means the participant consistency by CTT is in satisfactory category. Then, the INFIT MNSQ value is 1.00 with 0.18 SD or in range of acceptance of the IRT Rasch Model. The outcome of item difficulty may be viewed by adjusting proximation using the Quest.exe application for CTT, first at percent (%) in list categories and then at total number of right answers. See the following figure:

Figure 3.
Item Analysis Result for Item Difficulty

```
Item    1: item 1                                  Infit MNSQ =   .90
                                                      Disc =   .43

Categories          A         B         C         D*      missing

Count               7         3         2         34            0
Percent (%)      15.2       6.5       4.3       73.9
Pt-Biserial      -.30      -.13      -.24       .43
p-value          .023      .189      .056      .002
Mean Ability     -.03       .07      -.35       .70           NA

Step Labels                  1

Thresholds                -.66
Error                      .35
```

In the figure, there are 34 correct responses to item 1 with 73.9% on value. The conversion of the percentage value is at range 0.74 in difficulty index or in easy category. Nevertheless, for the IRT Rasch model, the thresholds result is indicating the item level of difficulty. See the figure below:

Figure 4.
Item Estimate (Thresholds)



	ITEM NAME	SCORE	MAXSCR	THRSH 1	INFT MNSQ	OUTFT MNSQ	INFT t	OUTFT t
1	item 1	34	46	-.66	.90	.78	-.6	-.7
				.35				

The whole correct answer for the same item is 34/46. The thresholds value of -0.66 indicates that the item is initially simple. As a consequence, both the CTT and IRT results estimated the question to be at an easy level for test takers. The following table summarizes the general findings of this research:

Table 4.
Item Difficulties

Items	CT		IRT	
	Indexing	Categories	Measurement Value (Thresholds)	Categories
1.	0.74	Easy	-0.66	Easy
2.	0.84	Easy	-1.20	Very Easy
3.	0.76	Easy	-0.78	Easy
4.	0.72	Easy	-0.54	Easy
5.	0.63	Moderate	-0.11	Easy
6.	0.65	Moderate	-0.22	Easy
7.	0.16	Difficult	2.34	Very Difficult
8.	0.41	Moderate	0.86	Difficult
9.	0.52	Moderate	0.38	Difficult
10.	0.39	Moderate	0.96	Difficult
11.	0.78	Easy	-0.91	Easy
12.	0.80	Easy	-1.05	Very Easy
13.	0.39	Moderate	0.96	Difficult
14.	0.83	Easy	-1.20	Very Easy
15.	0.65	Moderate	0.22	Difficult
16.	0.24	Difficult	1.75	Very Difficult
17.	0.87	Easy	-1.54	Very Easy
18.	0.70	Easy	-0.43	Easy
19.	0.47	Moderate	0.64	Difficult
20.	0.63	Moderate	-0.11	Easy
21.	0.62	Moderate	-0.07	Easy
22.	0.62	Moderate	-0.07	Easy
23.	0.29	Difficult	1.39	Very Difficult
24.	0.78	Easy	-0.87	Easy
25.	0.58	Moderate	0.15	Difficult
26.	0.87	Easy	-1.54	Very Difficult
27.	0.50	Moderate	0.54	Difficult
28.	0.41	Moderate	0.89	Difficult

Items	CT		IRT	
	Indexing	Categories	Measurement Value (Thresholds)	Categories
29.	0.28	Difficult	1.55	Very Difficult
30.	0.78	Easy	-0.89	Easy

The table above compares the difficulty level of items 1 to 30 in relation to both concepts. The CTT shows that the entire set of items is based on three levels of conceptual comprehension from the participants. The items 1 to 6 are indexed easy on an estimation range of 0.63 to 0.83. The items 8 to 10 are thereafter on a moderate level, with 0.41, 0.52, and 0.39 in sequential indexical order. Furthermore, item 7 is challenging since the index value is just 0.16 or lower than in CTT's indexing provision. Overall, there are 19 easy items, 8 moderate items, and 3 difficult ones.

Meanwhile, the threshold values in logit units represent the measurement of item difficulty level using IRT Rasch model. Very easy items have a measurement value less than -1, for example, items 12 and 14 with logit values of -1.05 and - 1.20 are classified as very easy. Items with values spanning -1 to 0 were classified as easy. For example, items 1 through 6 have logit values of -0.66, -1.20, -0.78, -0.54, -0.11, and - 0.22. Items with measurement values 0 to 1 are thus challenging, as are items 8 to 10 with threshold values of 0.86, 0.38, and 0.96 logit. The final and most difficult item has a measurement value greater than one logit, such as item 7 with a threshold value of 2.34 logit.

Figure 5.
Graphic of Item Estimates (Thresholds)

Figure 5 depicts the item's difficulty level as determined by the IRT Rasch Model with logit status ranging from 3.0 to -2.0. Each x symbolizes a single student. Item 7 on positive logit value +2.0 is the most difficult, whereas items 17 and 26 on negative logit value less than -1.0 are the easiest. As a result, the most difficult item implies a small amount of expected information gained from the item, whereas the easiest item implies a lower amount of expected information obtained.

Then, the discrimination power result is in the table below:

Table 5.
Discrimination Power

Item	Disc. Index	Categories	Interpretation	Items	Disc. Index	Categories	Interpretation
1.	0.43	High	Accepted	16.	0.32	Moderate	Accepted
2.	0.27	Moderate	Accepted	17.	0.23	Moderate	Accepted
3.	0.26	Moderate	Accepted	18.	0.45	High	Accepted
4.	0.40	Moderate	Accepted	19.	0.22	Moderate	Accepted
5.	0.25	Moderate	Accepted	20.	0.16	Low	Rejected
6.	0.33	Moderate	Accepted	21.	0.27	Moderate	Accepted
7.	0.32	Moderate	Accepted	22.	0.28	Moderate	Accepted
8.	0.27	Moderate	Accepted	23.	0.41	High	Accepted
9.	0.22	Moderate	Accepted	24.	0.22	Moderate	Accepted
10.	0.44	High	Accepted	25.	0.38	Moderate	Accepted
11.	0.35	Moderate	Accepted	26.	0.30	Moderate	Accepted
12.	0.41	High	Accepted	27.	0.24	Moderate	Accepted
13.	0.27	Moderate	Accepted	28.	0.33	Moderate	Accepted
14.	0.22	Moderate	Accepted	29.	0.51	High	Accepted
15.	0.30	Moderate	Accepted	30.	0.32	Moderate	Accepted

In table 7, 29 items are adequate for a test, however, item 20 index (0.16) shows a rejection in low category, therefore this item is unable to distinguish skills of students. The Quest.exe result for discrimination power following to CTT can be seen from the Pt-Biserial on list categories. See the representative result in followings.

Figure 6.
Item Analysis Result for Observed Response Item Discrimination Power

Item 29: item 29 — Infit MNSQ = .85, Disc = .51

Categories	A	B	C*	D	missing
Count	8	20	12	3	3
Percent (%)	18.6	46.5	27.9	7.0	
Pt-Biserial	-.09	-.31	.51	-.15	
p-value	.286	.020	.000	.177	
Mean Ability	.37	.25	1.20	.08	.14
Step Labels		1			
Thresholds		1.55			
Error		.36			

The Pt-Biserial for item 29 in figure above is 0.51, this result shows a high standard in comparing participants based on CTT indexing, owing to the result, this item is fit to use for a test or in high category. However, in Rasch Model with the same application, the discrimination power is critically analysing the individual abilities of students and items accordingly. It means to compare those who are capable to answer the item correctly with those who are not. The Rasch Model indexing may be demonstrated in the probability to reliability process through INFIT MNSQ and OUTFIT outcomes. See the figures below:

Figure 7.
Case Estimates in Input Order

NAME	SCORE	MAXSCR	ESTIMATE	ERROR	INFIT MNSQ	OUTFT MNSQ	INFT t	OUTFT t
1 001	16	29	.20	.41	1.02	1.00	.18	.09
2 002	22	30	1.21	.45	.85	.69	-.61	-.74

Figure 8.
Item Estimates (Thresholds) in Input Order

	ITEM NAME	SCORE	MAXSCR	THRSH 1	INFT MNSQ	OUTFT MNSQ	INFT t	OUTFT t
1	item 1	34	46	-.66	.90	.78	-.6	-.7
				.35				
2	item 2	38	46	-1.20	.98	.89	.0	-.1

Following to figure 6, the INFIT MNSQ of participant 001 and 002 are 1.02 and 0.85 respectively. The result indicates that they are adequate for the test given, however, the participant 002 is more compatible to respond the test than the participant 001, indicated by the ability value 1.21 and 0.20. Furthermore, in figure 7, the item 1 and 2 in INFIT

MNSQ value (0.90 and 0.98) show both items are applicable for all participants, which each item is answered correctly by 34 and 38 out of 46 participants.

Not only the questions, the quality of distraction is also fundamental in item analysis. Theoretically, a decent distractor is ideally chosen by minimally 5% of the test takers. Therefore, lower than that index means the item distractors must be revised or changed. The result of the item distractor for items in this research can be seen in the following table.

Table 6.
Distractor Effectiveness

Items	Distractor (%)				Description	Items	Distractor (%)				Description
	A	B	C	D			A	B	C	D	
1.	15.2	6.5	4.3	73.9	C – Revise	16.	10.9	21.7	43.5	23.9	Effective
2.	6.5	6.5	4.3	82.6	C – Revise	17.	4.3	87	4.3	4.3	A, C, D – Revise
3.	15.2	8.7	76.1	0	D – Change	18.	69.6	13	8.7	8.7	Effective
4.	8.7	71.7	0	19.6	C – Change	19.	40	46.7	6.7	6.7	Effective
5.	63	13	15.2	8.7	Effective	20.	6.5	19.6	63	10.9	Effective
6.	8.7	65.2	6.5	19.6	Effective	21.	6.7	22.2	62.2	8.9	Effective
7.	6.7	6.7	71.1	15.6	Effective	22.	62.2	2.2	26.7	8.9	B – Revise
8.	13	41.3	28.3	17.4	Effective	23.	6.5	26.1	37	30.4	Effective
9.	10.9	4.3	32.6	52.2	B – Revise	24.	77.8	4.4	6.7	11.1	B – Revise
10.	39.1	39.1	10.9	10.9	Effective	25.	57.8	4.4	35.6	2.2	B, D – Revise
11.	6.5	10.9	78.3	4.3	D – Revise	26.	87	8.7	2.2	2.2	C, D – Revise
12.	13	80.4	2.2	4.3	C, D – Revise	27.	9.5	7.1	50	33.3	Effective
13.	39.1	10.9	43.5	6.5	Effective	28.	6.8	9.1	43.2	40.9	Effective
14.	4.3	4.3	8.7	82.6	A, B – Revise	29.	18.6	46.5	27.9	7.0	Effective
15.	13	62.5	13	8.7	Effective	30.	6.7	6.7	8.9	77.8	Effective

The table 8 depicts of how effective the item distractors for the test. Of the 30 items, there are 17 items that the distractors worked effectively, while, there are still some distractors need adjustments; either small revision or change for more qualified distractors.

DISCUSSION

The item analysis is a significant phase in teaching and learning to specify inferences on the learners own learning output as well as to diagnose the compatibility of prepared and taught materials or the test items (Mahmoodi-shahrebabaki, 2018). Zuriyati (2016, as cited in Fitrianawati, 2017), furthermore, emphasizes the objectives of item analysis are to determine whether or not the items are compatible for learners, to improve the quality of items through three characteristics: difficulty level,

discrimination, and distractor appropriateness, to increase the validity and reliability of the items, and to revise irrelevant materials being taught based on how many correct responses learners come with. Owing to those characteristics, this research concern to measure the teacher's fabricated questions for 8th grade students based on CTT and IRT Rasch Model Analysis with QUEST.exe.

The first, the reliability is particularly to show how consistent the outcome of the items and response of persons are (Goolamally, 2019). An item is considered reliable if the statistical value is between 0.40 and 1.0 according to CTT and greater than 0.70 according to the IRT Rasch Model (Erfan et al., 2020; Goolamally, 2019; Heppi Yuslita, 2016; Susdelina, 2018). The evaluated item reliability in this research is 0.88 by CTT and 1.0 in INFIT MNSQ by IRT, making it suitable for testing students. Furthermore, the reliability, on small case estimation or person, demonstrated that test takers were consistent in their responses to the given items. The one above has a CTT value of 0.69 and an IRT value of 1.00 in INFIT MNSQ.

The second factor is determining the item level of difficulty, which is important for the assessment. Basically, the difficulty level is determined by the percentage of participants who properly answer the questions (Ashraf & Author, 2020). According to the range used, the items are regarded acceptable for the test takers of this research, with 19, 8, and 3 items being easy in index > 0.70, moderate in index 0.30 to 0.70, and difficult in index 3.0 by CTT. In addition, according to IRT, this research result confirms 6 very easy, 12 easy, 8 difficult, and 4 very difficult items, with the threshold values for each category being -1 for very easy, -1 to 0 for easy, 0 to 1 for tough, and > 1 for very difficult items (Susdelina, 2018), the detail values can be seen in table 6.

Then, the discrimination power of an item is the ability of item to compare participants who have high level of comprehension on materials to participants who has less understanding. To check the appropriateness of discrimination power, the CTT has suggested an acceptable range at minimally 0.20 or on moderate category (Susdelina, 2018), while in IRT, it must be in range of 0.77 to 1.33 INFIT MNSQ scores (Erfan et al., 2020; Goolamally, 2019; Susdelina, 2018). The statistical range of the discrimination results are given in table 7 that 29 items are sufficient as tests, however, 1 item requires a change. A brief comparison between CTT and IRT reveals that the IRT is not only assessing the item's usefulness as a test, but also determining students' capacity to respond to the item on a single to overall test performance (Erfan et al., 2020; Goolamally,

2019; Susdelina, 2018). Considering this statement, the details have been exemplified in figure 7 and 8 in case estimates and item thresholds in input orders.

Finally, the distractor effectiveness is also required to analyze when constructing a multiple-choice test format. Distractors are the wrong answer in this test type. The theory of CTT and IRT emphasized that the ideal distractor should have been decided by minimally 5% of the total students. When a distractor does not hit the range, the test maker needs to adjust the option. In this research, there are 17 items that the distractors work effectively, while 13 distractors of the items need either a revision or change for a qualified result. The detail for the distractor quantification is shown in table 8.

CONCLUSION

Item analysis is crucial in the teaching and learning process, offering valuable insights into exam quality and student comprehension. It evaluates test appropriateness, identifies item weaknesses, and allows for improvements through revision, modification, or removal. Keywords include item reliability (above 0.88 CTT or 1.0 IRT), difficulty level (easy > 0.70, difficult < 0.30 in CTT, or very easy < -1, very difficult > +1 in IRT), discrimination power (above 0.20), and distractor effectiveness (chosen by at least 5% of participants). This study demonstrated the utility of Quest.exe for item analysis but still faced limitations; such as a small sample size, a limited number of items, and reliance on fixed indices. Future research should address these gaps by using more specific models, larger and more diverse participant groups, and additional test formats, offering opportunities for teachers and material developers to enhance their materials effectively.

REFERENCES

Ashraf, Z. A., & Author, C. (2020). Classical and Modern Methods in Item Analysis of Test Tools Assistant Professor of Clinical Psychology, IMHANS, Kozhikode. *International Journal of Research and Review (Ijrrjournal.Com)*, *7*(5), 5.

Asrul, Ananda, R., & Rosinta. (2014). Evaluasi Pembajalaran. In *Ciptapustaka Media*.

Creswell, J. W. (2002). *Educational Research. Planning, Conducting, and Evaluating Quantitative and Qualitative Research*. Pearson Education Limited.

Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating*

quantitative and qualitative research (4rd Ed). MA: Pearson.

Danuwijaya, A. A. (2018). Item Analysis of Reading Comprehension Test for Post-Graduate Students. *English Review: Journal of English Education*, *7*(1), 29. https://doi.org/10.25134/erjee.v7i1.1493

Erfan, M., Maulyda, M. A., Hidayati, V. R., Astria, F. P., & Ratu, T. (2020). Tes Klasik Dan Model Rasch. *Indonesian Journal of Educational Research and Review*, *3*(1), 11–19.

Fitrianawati, M. (2017). Peran Analisis Butir Soal Guna Meningkatkan Kualitas Butir Soal, Kompetensi Guru, dan Hasil Belajar Peserta Didik. *Prosiding Seminar Nasional Pendidikan PGSD UMS & HDPGSDI Wilayah Jawa*, 282–295. https://publikasiilmiah.ums.ac.id/handle/11617/9117

Goolamally, S. H. C. & N. (2019). Validation of a two-tier multiple choice (2tmc) diagnostic instrument on the mole concept and solution concentration: a rasch analysis. *International Conference on Education*, *April*, 227–237.

Hartati, N., & Yogi, H. P. S. (2019). Item Analysis for a Better Quality Test. *English Language in Focus (ELIF)*, *2*(1), 59. https://doi.org/10.24853/elif.2.1.59-70

Heppi Yuslita, et. a. (2016). Analisis Tingkat Kesukaran Soal dan Daya Pembeda Soal Mata Ppelajaran Sejarah Kelas XI Semester Ganjil di SMA Negeri 5 Banda Aceh Tahun Pelajaran 2015-2016. *Jurnal Ilmiah Mahasisw Pendidikan Sejarah*, *1*(1), 131–138.

Jinyan Huang, Tiantian Shu, Yaxin Dong, D. Z. (2023). Constructing and Validating a Self-Assessment Scale for Chinese College English-Major Students' Feedback Knowledge Repertoire in EFL Academic Writing: Item Response Theory and Factor Analysis Approaches. *Assessing Writing*, *56*(100716). https://doi.org/https://doi.org/10.1016/j.asw.2023.100716

Karim, S. A., Sudiro, S., & Sakinah, S. (2021). Utilizing test items analysis to examine the level of difficulty and discriminating power in a teacher-made test. *EduLite: Journal of English Education, Literature and Culture*, *6*(2), 256. https://doi.org/10.30659/e.6.2.256-269

Khairuddinalfath, L. U. F. &. (2019). Analisis Kesukaran Soal, Daya Pembeda, dan Fungsi Distraktor. *Jurnal Komunikasi Dan Pendidikan Islam*, *8*(2), 37–64.

Law of the Republic of Indonesia Number 20. (2003). Act of the Republic of Indonesia on National Education System 1. *System*, *20*, 1–58.

Mahmoodi-shahrebabaki, M. (2018). *Assessment, Evaluation, and Testing: What are the Differences? February*, 3–5. https://www.researchgate.net/publication/323218570_Assessment_Evalu

ation_and_Testing_What_are_the_Differences

Ningsih, N. A., & Widowati, W. (2021). Utilizing Test Item Analysis to Portray the Quality of English Final Test. *English Teaching Journal : A Journal of English Literature, Language and Education*, *9*(2), 143. https://doi.org/10.25273/etj.v9i2.10867

Siri, A., & Freddano, M. (2011). The use of item analysis for the improvement of objective examinations. *Procedia - Social and Behavioral Sciences*, *29*, 188–197. https://doi.org/10.1016/j.sbspro.2011.11.224

Suarga, S. (2019). Hakikat, Tujuan Dan Fungsi Evaluasi Dalam Pengembangan Pembelajaran. *Inspiratif Pendidikan*, *8*(1), 327–338. https://doi.org/10.24252/ip.v8i1.7844

Suek, L. A. (2021). Item Analysis of an English Summative Test. *PEJLaC: Pattimura Excellence Journal of Language and Culture*, *1*(1), 9–18. https://doi.org/10.30598/pejlac.v1.i1.pp9-18

Susdelina, E. a. (2018). Analisis Kualitas Instrumen Pengukuran Pemahaman Konsep Persamaan Kuadrat Melalui Teori Tes Klasik Dan Rasch Model. *Jurnal Kiprah*, *6*(1), 41–48. https://doi.org/10.31629/kiprah.v6i1.574