REID (Research and Evaluation in Education), 9(1), 2023, 24-36
Available online at: http://journal.uny.ac.id/index.php/reid

Item analysis of reading comprehension questions for English proficiency test using Rasch model

Henda Harmantia Dewi[1]*; Siti Maftuhah Damio[2]; Sukarno[1]
[1]Universitas Negeri Yogyakarta, Indonesia
[2]Universiti Teknologi MARA, Malaysia
*Corresponding Author. E-mail: hendaharmantia.2021@student.uny.ac.id

ARTICLE INFO

Article History
Submitted:
26 September 2022
Revised:
12 December 2022
Accepted:
19 May 2023

Keywords
reading comprehension; English proficiency test; item analysis; Rasch model

Scan Me:

ABSTRACT

The need to take English as a foreign language proficiency test (known as TOEFL [Test of English Language Proficiency]) has been gaining popularity in Indonesia. The increasing demands for such a test and its expensive cost have reinforced many institutions to develop TOEFL instruments and administer the test internally. However, constructing a test instrument is a complex process that makes conducting item analysis become more challenging. Meanwhile, item analysis is crucial to assess the items' quality. Therefore, this study reported the results of statistically analyzing 20 questions of TOEFL reading comprehension that were analyzed in terms of the test reliability, the item and person fit, and the items' difficulty level. Thirty-eight members of the English Department Students' Association of a state university in West Java participated in this study by taking the reading test. The data were analyzed using the Rasch model by utilizing the Quest program. The results showed that four items (36.8%) did not fulfill the ideal criteria of a valid test because they were too easy and too difficult to be given to the target test takers; thus, they needed to be discarded. Meanwhile, 16 items (63.2%) are of good quality and can be used immediately in the proficiency test, especially to measure reading comprehension skills, because they have fulfilled the standard requirements for a valid test. The findings have provided insight into the importance of item analysis in validating test instruments to improve the test quality for future administrations.

To cite this article (in APA style):
Dewi, H., Damio, S., & Sukarno, S. (2023). Item analysis of reading comprehension questions for English proficiency test using Rasch model. *REID (Research and Evaluation in Education), 9*(1), 24-36. doi:https://doi.org/10.21831/reid.v9i1.53514

INTRODUCTION

English language proficiency has been viewed as structural, communicative, social, and functional competencies that put the language use, language users, contexts, situations, or interactants into account (Mouvet & Taverniers, 2022). It has also fundamental tenets with the everyday practice of language teaching as it can be used as an evaluative criterion, focusing on particular selected aspects, to assess learning outcomes. It is often associated with "the notions of measurement, accuracy, reliability, and trustworthiness" (Leung, 2022, p. 56). In a more modern definition, it has been deemed an essential requirement for academic success, especially at the university level, because it shows learners' capacities of language acquisition that can be used in educational, institutional, and professional contexts (Mouvet & Taverniers, 2022). It is frequently included in the admission purposes, such as in higher education in Singapore, Belgium, and South Korea (Bo et al., 2022; Choi, 2008; Mouvet & Taverniers, 2022) since the results can show learners' corresponding English levels. Learners' proficiency can be further predicted through internationally standardized tests in which the benchmark has been set, for instance, by the International

ISSN: 2460-6995 (Online)

English Language Testing System (IELTS), the Test of English as a Foreign Language (TOEFL), or the Common European Framework of Reference for Languages (CEFR). In addition, a locally developed test can also be an alternative because it aligns with the local institution's curricula (Bo et al., 2022).

Similar to other countries, the need to take the English as a foreign language proficiency test, especially TOEFL, has also been gaining popularity in Indonesia. Many institutions such as colleges, universities, and language programs have been administering such a test for many different purposes such as to meet university and graduation requirements, obtain scholarships, and apply for a job (Danuwijaya, 2018; Golubovich et al., 2018; Karjo & Ronaldo, 2019; Renandya et al., 2018; Suryani & Khadijah, 2021). These purposes align with the official TOEFL ITP (Institutional Testing Program) objectives in which the test is used for placement, monitoring progress, evaluation, exit testing, program admission, and many more (ETS TOEFL ITP, 2022). Besides, TOEFL is an official test that has been administered around the globe, such as in Japan, Mexico, and 50 more countries including Indonesia, by the Educational Testing Service (ETS) to ensure the credibility of the results (ETS TOEFL ITP, 2022; Golubovich et al., 2018). Because of these reasons, TOEFL has become one of the popular proficiency tests that are widely used in the Indonesian context.

Despite its objectives and popularity, the cost and practicality of the official TOEFL test have become an issue as it is cost-prohibitive (Golubovich et al., 2018; Mustafa & Apriadi, 2014). Both test takers and institutions require inexpensive tests that can show the results of the test takers' English proficiency as credible as the ones from the TOEFL test. Consequently, many institutions have developed TOEFL instruments and administered the test locally for internal purposes (Danuwijaya, 2018; Mustafa & Apriadi, 2014). Constructing a test instrument, however, is a complex process that requires planning, test preparation and administration, statistical analysis, and test result reports (Downing, 2010). The test items should also be constructed with a question format that is similar to the ones in the actual TOEFL test by considering the text selection, the standardized multiple-choice questions (MCQs), and the like (Mustafa & Apriadi, 2014). These complexities have resulted in a more challenging process of piloting and analyzing the test items. Meanwhile, item analysis is a crucial stage for instrument developers to see the test items' effectiveness and judge their quality (Brown, 2012; Moses, 2017) before they are publicly administered.

Existing research recognizes the critical role that is played by item analysis. Data from several studies suggest that the analyzed items commonly emerged from summative tests in school contexts. This has been seen in the case of item development for English summative tests and English try out (Jannah et al., 2021; Maharani & Putro, 2020; Sugianto, 2020; Yumelking, 2019) as well as for the final semester exam of the Indonesian subject (Azizah et al., 2022) at junior and senior high school levels. Those tests comprised vocabulary, grammar, and reading skills. A study focusing on a reading comprehension test for a reading supplementary course (Saswati, 2021) was also found at the university level. These results indicate that item analysis has been rarely used for other purposes, but mostly for analyzing school test instruments, whereas TOEFL questions development that was mentioned to be highly demanded has remained scarcely investigated.

The scarcity was further justified by only a few studies reporting the results of item analysis for TOEFL instruments, such as reading comprehension (Danuwijaya, 2018; Mustafa & Apriadi, 2014), listening skills (Sacko & Haidara, 2018), as well as the structure and written expression (Mustafa, 2015; Thu, 2019). Although the literature shows that item analysis for TOEFL has been previously examined, there currently needs to be more research concerning its practice. Therefore, this study will focus on statistically analyzing the questions of one of the skills, reading comprehension. It is evident that the questions for reading were always included in many types of test instruments. Moreover, reading in TOEFL is also one of the most important skills whose results can reflect the test takers' understanding of English written information not only by under-

standing separate letters or words (bottom-up processes) but also the whole information (top-down processes) (Brown & Abeywickrama, 2018). Therefore, this study focused on reporting the results of analyzing reading comprehension questions for TOEFL purposes. Although the test usually comprises 50 questions (ETS TOEFL ITP, 2022), due to the time constraint, only 20 questions were developed.

Furthermore, many existing studies on item analysis have used many different perspectives such as classical test theory (CTT) and item response theory (IRT). Both theories utilized various software to analyze data such as the Laboratory of Educational Research Test Analysis Package (Lertap), Anbuso, Bilog, Winsteps, Mplus, IRTPRO, RUMM2030, Lisrel 8.8, and Quest program (Faradillah & Adlina, 2021; Hagquist & Andrich, 2017; Muchlisin et al., 2019; Ndayizeye, 2017; Rahim & Haryanto, 2021; Wahyuni & Kartowagiran, 2018). The CTT, however, has been long dropped because of its numerous flaws such as how the test takers' ability affects the item's parameter and how the results will depend on the item's characteristics (Muchlisin et al., 2019; Rahim & Haryanto, 2021). Hence, the present study used the IRT by utilizing the Rasch model because it can cover the shortfalls of the CTT. In Indonesia, the Rasch model has become increasingly necessary as a tool for analyzing data on assessment (Hayat et al., 2020), such as in English (Ndayizeye, 2017), Islamic (Wahyuni & Kartowagiran, 2018), Physics (Habibi et al., 2019), and even Mathematics Education (Isnani et al., 2019; Rizbudiani et al., 2021). It also has important features to measure three indicators namely the items' and person's reliability, the item's fit to the model, and the items' difficulty level. Thus, the results can be advantageous to assess the quality of the items because test items' validity and reliability are important to ensure the quality of test scores.

The aforementioned situations indicate an empirical gap in which too little attention has been paid to analyzing locally developed TOEFL reading questions, especially for the purpose of identifying the English language proficiency of Indonesian higher education students. Moreover, the study used a methodological gap as it utilized a more recent approach to the Rasch Model that can complement the flaws of the previous theory. The model has also been mainly applied in analyzing items for summative assessments, leaving the locally developed questions under-researched. Because of these reasons, the present study aims at providing institutions with validated reading TOEFL questions that can be used for internal purposes such as for conducting a proficiency test, or by test takers for TOEFL preparation. With the increased demands for the TOEFL test, the availability of qualified test items should be adequate to meet the needs. Therefore, it is necessary to develop and analyze the test items that can function as test item banks and be beneficial for future use.

METHOD

Test Item Specification

Since the study focused on developing reading comprehension questions for TOEFL purposes, academic reading was used as the genre of the test. The academic genre includes general interest articles, technical reports, reference materials, textbooks, theses, essays, papers, etc. (Brown & Abeywickrama, 2018). Based on this consideration, there were five text types used in developing the test items. They were a biography (J.K. Rowling), a descriptive text (Cybercrime), a hortatory exposition (Autism), a report (Bilingualism), and an explanation text (Drone). These topics were selected because they are commonly discussed in the academic context of different fields. The question types of the test were adapted mainly from the reading skills required in TOEFL ITP (Phillips, 2001) and from some of the skills expected in the iBT (Internet-based Test) TOEFL (ETS TOEFL, 2022). The decision to select the iBT format was to balance the difficulty level based on Bloom's Taxonomy. Table 1 presents the breakdown of the test items' specifications for the 20 MCQs according to their question type.

Table 1. Test Item Specification

Question Type	Numbers of Questions	Item Numbers
Finding the main idea	1	10
Recognizing the organization of ideas	1	20
Answering stated detail questions	1	6
Finding unstated details	3	4, 14, 17
Finding pronoun reference	2	2, 13
Answering implied questions	2	3, 5
Answering transition questions	2	8, 16
Determining meaning (vocabulary questions)	3	1, 7, 11
Determining the tone, purpose, or course	4	9, 12, 15, 18
Simplifying a sentence	1	19
Total Questions	**20**	

Respondents and Pilot Testing

After the test had been developed, a peer-review process was conducted, and feedback was given by a graduate student at a state university in Yogyakarta. This process is a content validity test because it aims to evaluate the degree to which the test instrument evaluates all aspects to be assessed (Delgado-Rico et al., 2012), especially in the TOEFL test. There are at least three important elements in test instruments that have to be evaluated (Brown & Abeywickrama, 2018). The first includes the quality of the items (content) e.g., the clarity of instructions, the questions' level of difficulty, stem, key, and distractors. The second comprises vocabulary such as word choice. Moreover, the last one consists of grammar and mechanics e.g., grammar accuracy, capitalization, punctuation, and spelling. The results required minor changes in numbers 19 and 20 as they led to miscomprehension of the questions. Therefore, the changes in wording were made to ensure that test takers understood the questions when answering the items.

The instruments were then administered to former members of EDSA (English Department Students Association) at a state university in West Java, from the year 2016 to 2020. There were 38 members who willingly participated as respondents. Twenty-eight of them were female, while ten of them were male. The selection criterion was based on the respondents' characteristics such as (a) having basic knowledge of English, especially because they were majoring in English Education and (b) having been familiar with such a test because they had passed the EPT (English Proficiency Test) required by the university as the graduation requirement. Moreover, they were available and accessible at the time when the test was administered (Cohen et al., 2018). Thus, they were considered suitable candidates for pilot testing.

The test was administered via Google Form from May 29, 2022, to June 4, 2022 (seven days). The respondents were informed to take the test on their available time; complete the test by using a laptop, an iPad, or a device that has a big screen to help them read the questions effectively; and answer the questions honestly without using any search engines and dictionaries. However, since the test was conducted remotely and individually, the process of undertaking the test could not be monitored. Thus, the only option was to fully rely on the data obtained from the form.

Data Analysis

Rasch model, a model developed from the IRT, was employed to analyze the test items. It was used because of its dichotomous scoring system consisting of two categories namely the correct answer with a score of 1 and the incorrect answer with a score of 0 (Isnani et al., 2019). In this model, item and person data are the basic parameters that are used for estimating true scores (Hagquist & Andrich, 2017). They can indicate the level of the test takers' ability and the degree of the items' difficulty. This advantage is one of the primary features of the model because it takes into account each test taker's responses regardless of the order of the items' difficulty level (Isnani et al., 2019). The parameters of discriminating power are also assumed to have been

the same for all of the test items (Hayat et al., 2020). Moreover, the Rasch model can fulfill five criteria for assessment in education. They include (a) producing linear measures, (b) overcoming missing data, (c) giving estimates of precision, (d) having devices for detecting misfits, and (e) separating the parameters of the object being measured and of the measurement instrument (Wright & Mok, 2004). Because of these advantages, the model is considered to be more accurate than any other existing one as well as the most objective measurement model.

To help analyze the data, the Quest program was utilized. Quest is an interactive test analysis system published by the Australian Council for Educational Research Limited (ACER) (Izard, 2005). The central elements in the QUEST program are IRTs that have been adjusted to the Rasch model (Habibi et al., 2019). It can be used to analyze MCQs from the Rasch perspective including three indicators. Firstly, the test reliability can be obtained from the item and case estimates of the Quest output. The overall fit of items and person was based on the average value of the INFIT Mean of Square (INFIT MNSQ) and the standard deviation or the average value of the INFIT Mean INFIT *t*. Secondly, the separate item and person compatibility to the model can be acquired from the INFIT MNSQ value or the INFIT *t* value (Rizbudiani et al., 2021). Thirdly, the items' difficulty level can be located from the threshold value of the program output (Setyawarno, 2017). This difficulty level can also be used to describe the relationship between the test items and the person.

Investigator triangulation was utilized to contribute to the internal validity of analyzing and interpreting the data. Since each investigator has different and complementary skills, it can reduce the possibility of obtaining inaccurate and inconsistent results, then increase the credibility of the findings (Thurmond, 2001). In this study, the initial data gathering, and analysis processes were conducted and then interpreted by the first author. Afterward, the second author reevaluated the data display and the data interpretation to be well connected and clarified. Finally, the third author reassessed the data analysis and interpretation more holistically to ensure that the results were properly obtained by taking the Rasch Model theory into account.

FINDINGS AND DISCUSSION

Test Reliability

Reliability evaluation aims to identify whether a measuring instrument is highly or poorly performing (Hamon & Mesbah, 2002). To analyze the test reliability, the Quest program provides reliability values that can be obtained from the reliability of item and case estimates. These estimates can show the degree of consistency of the instrument (Ofianto, 2018) and the error in the measurement process because it evaluates the probability of the test's stability when it is administered at different times to the same individuals or using the same standards (Faradillah & Adlina, 2021; Kimberlin & Winterstein, 2008). The results of the statistical analysis are summarized in Table 2 and the reliability criteria based on the Rasch model (Setyawarno, 2017) are presented in Table 3.

Table 2. Statistical Summary of Item and Person Estimates

Estimates	Mean	SD	SD (adj)	Reliability	Infit Mean Square		Outfit Mean Square		Infit *t*		Outfit *t*	
					Mean	SD	Mean	SD	Mean	SD	Mean	SD
Item	.00	1.37	1.30	.89	1.00	.07	1.02	.15	.04	.58	.12	.53
Case	-.01	.62	.32	.26	.99	.21	1.02	.54	-.02	1.01	.14	.77

Table 2 illustrates that the items' reliability level is .89. Based on the criteria in Table 3, the reliability value of .89 can be interpreted as good. The higher the value, the more reliable the quality of the items. The findings from the overall item analysis also show that the items fit the

Rasch model because the INFIT MNSQ value is between the standard range of 0.77 – 1.33. This compatibility is consistent with the analysis result of the OUTFIT *t*, revealing that the items conform to the model because the OUTFIT t value is less than 2.00. Thus far, it concludes that the overall items have high reliability and fit the proposed model. Therefore, the test items can measure what they intend to measure, in this case the English proficiency of reading comprehension.

Table 3. Item and Person Reliability in Rasch Model

Fit Indices	Interpretation	Infit MNSQ	Interpretation	Outfit *t*	Interpretation
<0.67	Low	>1.33	Misfit	≤ 2.00	Fit
0.67-0.80	Sufficient	0.77-1.33	Fit	≥ 2.00	Misfit
0.81-0.90	Good	<0.77	Misfit		
0.91-0.94	Very Good				
>0.94	Excellent				

On the other hand, the reliability of case estimates is only .26. This result shows that the test takers do not give consistent results. The low person's reliability also signifies that the same results will not probably be obtained if the test is re-administered to different test takers. A likely explanation for these findings is because of the test takers' attempt answering the questions negligently (Ardiyanti, 2016). Their negligent answers resulted in inconsistent and unnatural response patterns (Faradillah & Febriani, 2021). Besides, the low reliability is also likely because the number of the test takers was less than 100, but only 38. This finding is in accord with previous studies conducted by Lia et al. (2020) as well as Faradillah and Adlina (2021). They obtained high person reliability because they administered 35 and 51 items to more than 100 respondents, that was 130 and 245 test takers respectively. Moreover, another study supported this finding because low reliability was achieved from the administration of 30 questions to 65 students, in which the participants were lower than 100 (Pratama, 2020). Thus, it can be inferred that the more the number of test takers, the higher the person's reliability value.

In conclusion, the overall item is confirmed to be reliable, good, and has met the basic requirements of fitting the Rasch model. However, the outcome of the case reliability is contrary to the item estimates because the quality is revealed to be low. Despite this low quality, the analysis result verifies that the overall participants still fit the proposed model.

Estimation of Person Fit

After having discussed the estimates of the overall item and case reliability, this subsection will address the estimation of each fit person. This estimation is beneficial to provide findings that can be either supportive or contradictory to the proceeding finding. To highlight, the Rasch model analysis allows the opportunity of analyzing person fit from the aspect of the unusual test takers' response patterns (Faradillah & Febriani, 2021). Table 4 illustrates the analysis results of the person fitting the model on an individual basis that are analyzed from the criteria of INFIT MNSQ and also INFIT *t* (Habibi et al., 2019). The criteria specification is presented under Table 4.

As shown in Table 4, a more detailed breakdown of each of the test takers who meet and do not meet the given criteria is presented. It is apparent from the table that 36 out of 38 respondents correspond with the provided criteria. This result also represents almost the entire population, with a percentage of 94.7%. Nevertheless, two out of 38 respondents (5.3%) do not match the requirements for a person who fits the model. Despite the fact that a misfit person is in existence, its percentage does not exceed the fit one. Therefore, in this study, it is confirmed that the results of each person fit support the findings that are obtained in the overall person fit in the previous section. It is prevalent that the respondents generally fit the projected model.

Table 4. Person Fitting the Rasch Model

Items	Infit MNSQ	Infit *t*	Criterion	Items	Infit MNSQ	Infit *t*	Criterion	Items	Infit MNSQ	Infit *t*	Criterion
1	.84	-.24	Fit	14	.98	-.06	Fit	27	.75	-1.23	Fit
2	.83	-.83	Fit	15	.86	-.70	Fit	28	1.16	.63	Fit
3	.81	-.98	Fit	16	.70	-1.70	Fit	29	1.54	2.40	Misfit
4	.98	-.03	Fit	17	1.11	.56	Fit	30	.78	-1.04	Fit
5	.68	-1.79	Fit	18	.92	-.33	Fit	31	1.01	.14	Fit
6	1.24	1.19	Fit	19	.86	-.62	Fit	32	.81	-.79	Fit
7	1.04	.24	Fit	20	1.13	.60	Fit	33	1.21	1.05	Fit
8	1.53	2.41	Misfit	21	1.22	1.12	Fit	34	.94	-.25	Fit
9	1.11	.58	Fit	22	.83	-.86	Fit	35	1.16	.63	Fit
10	.81	-.95	Fit	23	1.03	.24	Fit	36	1.18	.94	Fit
11	1.13	.71	Fit	24	.76	-1.15	Fit	37	1.14	.74	Fit
12	.98	.02	Fit	25	.94	-.24	Fit	38	1.01	.12	Fit
13	.69	-1.71	Fit	26	1.11	.43	Fit				

Criteria for fit person: 0, 77 ≤ Infit MNSQ ≤ 1, 33 OR -2 ≤ Infit *t* ≤ 2

Estimation of Item Fit

Item fit can show that the item functions properly based on certain criteria. The recommended criteria are identical to the one in the person fit in the preceding subsection. Similarly, the item fit will also be estimated singly. The output from the QUEST program can solicit information about the feasibility of each item based on the Rasch model. The interpretation of the analysis results can be concluded as follows: (1) if an item fits, it shows that it can work properly; and (2) if an item misfits, it shows the test takers' misconception of the item (Faradillah & Febriani, 2021). Table 5 illustrates the distribution of the 20 items based on their compatibility with the model.

Table 5. Items Fitting the Rasch Model

Items	Infit MNSQ	Infit *t*	Criterion	Items	Infit MNSQ	Infit *t*	Criterion
1	.98	-.1	Fit	11	1.02	.2	Fit
2	1.03	.2	Fit	12	1.09	.3	Fit
3	1.11	1.1	Fit	13	1.04	.4	Fit
4	.88	-1.1	Fit	14	.93	-.5	Fit
5	Perfect score		Misfit	15	1.07	.4	Fit
6	.99	-.9	Fit	16	.92	-.4	Fit
7	1.08	.8	Fit	17	1.08	.8	Fit
8	1.02	.2	Fit	18	.90	-.4	Fit
9	.93	-.6	Fit	19	1.00	.0	Fit
10	1.04	.3	Fit	20	1.02	.3	Fit

Criteria for fit items: 0, 77 ≤ Infit MNSQ ≤ 1, 33 OR -2 ≤ Infit *t* ≤ 2

The data in Table 5 reveal an unexpected result. Although the previous findings reported that the overall items conform to the Rasch model, further analysis confirms that only 95% of them fit the model. Meanwhile, the remaining 5% do not. One item, number 5, consists of a zero score due to the fact that all of the test takers could answer it perfectly. This result suggests that the item misfits the model especially because it is too easy and does not meet the validity criteria; thus, it needs to be excluded from the list. On the other hand, the remaining significant percentage of the item can be used to measure the test takers' reading comprehension because they pass the recommended criteria.

Item Difficulty Level

The items' difficulty level can be categorized into four: very difficult, difficult, moderate, easy, and very easy (Setyawarno, 2017). The threshold value of the Quest output serves as the difficulty index. Theoretically, the level can be analyzed from the range of minus infinity to infinity (Muchlisin et al., 2019). The categorization of all 20 items was calculated and classified into the criteria in Table 6. Meanwhile, Table 7 shows the summary of the percentage along with the items' quality judgment. To note, since item 5 does not fit the model, it has been removed from the list.

Table 6. Items' Difficulty Level

Items	Threshold	Interpretation	Items	Threshold	Interpretation
1	-.67	Moderate	11	.11	Moderate
2	-2.48	Very easy	12	2.48	Very difficult
3	.44	Moderate	13	-3.60	Very easy
4	-.44	Moderate	14	-.67	Moderate
5	-	-	15	1.20	Difficult
6	-.67	Moderate	16	1.06	Difficult
7	-.44	Moderate	17	.33	Moderate
8	1.20	Difficult	18	1.20	Difficult
9	-.44	Moderate	19	.33	Moderate
10	.92	Moderate	20	.11	Moderate
Criteria	b>2 Very difficult		-1≤b≤1 Moderate		-1≤b≥-2 Easy
	1<b≤2 Difficult				b<-2 Very easy

Table 7. Percentage of Items' Difficulty Level and Quality Judgment

Criteria	Quality	Frequency	Item Numbers	Percentage
Very difficult	Not good	1	12	5.3%
Difficult	Good	4	8, 15, 16, 18	21.1%
Moderate	Good	12	1, 3, 4, 6, 7, 9, 10, 11, 14, 17, 19, 20	63.2%
Easy	-	0		0%
Very easy	Not good	2	2, 13	10.5%

Table 6 reveals that the majority of the items, 12 in sum, are classified into the moderate level. These numbers represent more than half of the items, with a 63.2% percentage. The items that are categorized into this level include the following numbers: 1, 7, and 11 (determining meaning), 3 (answering implied questions), 4, 14, and 17 (finding unstated details), 6 (answering stated detail questions), 9 (determining the tone of the passage), 19 (simplifying a sentence), and 20 (recognizing the organization of ideas). This difficulty level is consistent with the criteria proposed by Setyawarno (2017) in which the threshold value of the questions ranges from -1≤b≤1. Furthermore, the data indicate that the difficult category places second. It consists of 4 test items including numbers 8 and 16 (answering transitioning questions) as well as 15 and 18 (determining the purpose and course of the passage). Table 6 and Table 7 show that a little over a fifth of the test items have a high difficulty index, ranging from 1<b≤2. This positive value indicates relatively more difficult items (Finch & French, 2015).

The most surprising aspect of the data is that none of the items is distributed into the easy level. Rather, two items drop straightforwardly into the very easy classification because the value is lower than -2. It means, 10.5% of the items are allocated into the very easy level comprising numbers 2 and 3 aspire to find pronoun references. Moreover, the least occurrence is presented by the very difficult classification with only one frequency of the question. This result also indicates that this level has the lowest percentage among all, 5.3%. The number of the item asso-

ciated with this level includes number 12 which aims to determine the passage's tone. Consistent with the literature, the item is classified as very difficult because the item's value falls between $1<b\leq2$.

From the tables, it can be interpreted that the items with a moderate level are in the majority. The difficult one is shown to be in the second position, following the moderate level. On the other hand, the very easy items have the second lowest percentage, proceding the very difficult category. To conclude, if organized from the items that have the highest to the smallest percentage, the order is as follows: moderate, difficult, very easy, and very difficult as none of the items is classified into the easy level.

Ideally, test items should not be too easy or too difficult for the intended test takers (Fulcher & Davidson, 2007). The recommended difficulty range is between 30%-70%, corresponding to the moderate category (Hingorjo & Jaleel, 2012). Kunandar (2013) also asserted that an ideal test consists of 25% easy, 50% moderate, and 25% difficult items. Therefore, the results of the analysis show that even though the easy and difficult items are not equally distributed, the test can still be considered ideal. As many as 63.2% of the moderate questions fall around the standard percentage of an ideal test of which this finding is in accordance with the assertion of Hingorjo and Jaleel (2012). The average level of difficulty should be integrated into the test because it can improve the test takers' scores (Hingorjo & Jaleel, 2012). The difficult ones, on the other hand, should be reviewed for "confusing language, areas of controversy, or even an incorrect key" because test takers tend to put a lot of guesswork into selecting the correct answer (Hingorjo & Jaleel, 2012, p. 143). After all, the easy items can be put at the beginning of the test as warming-up questions.

Furthermore, the test items were initially assumed to be organized from easy to difficult, following Bloom's Taxonomy. Yet, the ideal proportion of easy, moderate, and difficult questions does not indicate that the items are automatically organized based on the difficulty level. Thus, the order of the items should be re-organized by placing the easy ones in the beginning, followed by the moderate and difficult questions (Hingorjo & Jaleel, 2012). However, since the test focuses on reading skills that are arranged based on the type of texts, the organization should be arranged within the text types, then the test as a whole.

These results suggest that the quality judgment can now be decided. The item is regarded to be good if it has a difficulty value ranging from -2.0 to 2.0 (Muchlisin et al., 2019). Table 7 indicates that 84.2% of the items are classified as good quality items because the values of a little over four-fifths of the questions fall within the target criteria. On the other hand, 15.8% of the items are considered the opposite because the values lie outside of the target range. The 15.8% of the items need to be discarded because they cannot be used in the actual test.

CONCLUSION

Based on the results of the analysis carried out using the Rasch model analysis, one item does not fit the target model; thus it needs to be eliminated from the list. Moreover, three items need to be removed from the difficulty analysis because they are classified as low-quality items. It indicates that these four items (36.8%) do not fulfill the ideal criteria of a valid test because they are too easy and too difficult to be given to the target test takers. Furthermore, the results signify that 16 items (63.2%) are of good quality. Hence, they can be used immediately in the proficiency test because they have fulfilled the standard requirements for a valid test. They can be used for assessing test takers' English proficiency in terms of their reading comprehension skills.

These results indicate that developing test items requires a complex process from planning the items through the table of specifications, developing the items, selecting the appropriate text, administering the test, to analyzing and evaluating the results. Taken together, the results suggest two recommendations for future test development. Firstly, constructing valid and reliable proficiency tests calls for the involvement of professional and well-trained personnel who have different expertise from experts in the fields, test specialists, editors, and the like. Secondly, experi-

enced staff can be formed through the continual high-quality practice of item writing, bearing in mind the intended purposes. Therefore, the government, stakeholders, and/or policymakers have to facilitate test designers, and everyone is involved with workshops or training to construct valid and reliable proficiency tests. By doing so, test designers can ensure that the test can assess what it intends to assess. Last but not least, as this project involved a small scale of respondents, it might be possible for future researchers to analyze test-item developments on a bigger scale. Testing the data from different methodologies will also give broader results to the perspectives.

REFERENCES

Ardiyanti, D. (2016). Aplikasi model Rasch pada pengembangan skala efikasi diri dalam pengambilan keputusan karier siswa. *Jurnal Psikologi*, *43*(3), 248–263. https://doi.org/10.22146/jpsi.17801

Azizah, N., Suseno, M., & Hayat, B. (2022). Item analysis of the rasch model items in the final semester exam indonesian language lesson. *World Journal of English Language*, *12*(1), 15–26. https://doi.org/10.5430/wjel.v12n1p15

Bo, W. V., Fu, M., & Lim, W. Y. (2022). Revisiting English language proficiency and its impact on the academic performance of domestic university students in Singapore. *Language Testing, 40*(1). https://doi.org/10.1177/02655322211064629

Brown, H. D., & Abeywickrama, P. (2018). *Language assessment: Principles and classroom practices* (3rd Ed.). Pearson/Longman.

Brown, J. D. (2012). Classical test theory. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing*. Routledge.

Choi, I. C. (2008). The impact of EFL testing on EFL education in Korea. *Language Testing*, *25*(1), 39–62. https://doi.org/10.1177/0265532207083744

Cohen, L., Manion, L., & Morrison, K. (2018). *Research methods in education* (8th Ed.). Routledge.

Danuwijaya, A. A. (2018). Item analysis of reading comprehension test for post-graduate students. *English Review: Journal of English Education*, *7*(1), 29-40. https://doi.org/10.25134/erjee.v7i1.1493

Delgado-Rico, Carretero-Dios, H., & Ruch, W. (2012). Content validity evidences in test development: An applied perspective. *International Journal of Clinical and Health Psychology España*, *12*(3), 449–460. https://doi.org/10.5167/uzh-64551

Downing, S. M. (2010). Twelve steps for effective test development. In S. M. Downing & S. M. Downing (Eds.), *Handbook of test development*. Routledge.

ETS TOEFL. (2022). *TOEFL iBT® reading section*. ETS. https://www.ets.org/toefl/test-takers/ibt/about/content/reading/

ETS TOEFL ITP. (2022). *TOEFL ITP® assessment series*. ETS. https://www.ets.org/toefl_itp/

Faradillah, A., & Adlina, S. (2021). Validity of critical thinking skills instrument on prospective Mathematics teachers. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *25*(2), 126-137. https://doi.org/10.21831/pep.v25i2.40662

Faradillah, A., & Febriani, L. (2021). Mathematical trauma students' junior high school based on grade and gender. *Infinity Journal*, *10*(1), 53-67. https://doi.org/10.22460/infinity.v10i1.p53-68

Finch, W. H., & French, B. F. (2015). *Latent variable modeling with R* (W. H. Finch, Ed.; 1st Ed.). Taylor & Francis.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book* (1st Ed.). Routledge.

Golubovich, J., Tolentino, F., & Papageorgiou, S. (2018). Examining the applications and opinions of the TOEFL ITP® assessment series test scores in three countries. *ETS Research Report Series, 2018*(1), 1-30. https://doi.org/10.1002/ets2.12231

Habibi, H., Jumadi, J., & Mundilarto, M. (2019). The Rasch-rating scale model to identify learning difficulties of physics students based on self-regulation skills. *International Journal of Evaluation and Research in Education*, *8*(4), 659–665. https://doi.org/10.11591/ijere.v8i4.20292

Hagquist, C., & Andrich, D. (2017). Recent advances in analysis of differential item functioning in health research using the Rasch model. *Health and Quality of Life Outcomes*, *15*(1), 181. https://doi.org/10.1186/s12955-017-0755-0

Hamon, A., & Mesbah, M. (2002). Questionnaire reliability under the Rasch model. In *Statistical methods for quality of life studies* (pp. 155–168). Springer.

Hayat, B., Dwirifqi, M., Putra, K., & Suryadi, B. (2020). Comparing item parameter estimates and fit statistics of the Rasch model from three different traditions. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *24*(1), 39–50. https://doi.org/10.21831/pep.v24i1

Hingorjo, M. R., & Jaleel, F. (2012). Analysis of one-best MCQs: The difficulty index, discrimination index and distractor efficiency. *JPMA-Journal of the Pakistan Medical Association*, *62*(2), 142–147.

Isnani, I., Utami, W. B., Susongko, P., & Lestiani, H. T. (2019). Estimation of college students' ability on real analysis course using Rasch model. *REID (Research and Evaluation in Education)*, *5*(2), 95–102. https://doi.org/10.21831/reid.v5i2.20924

Izard, J. (2005). Trial testing and item analysis in test construction. In K. Ross (Ed.), *Quantitative research methods in educational planning*. UNESCO International Institute for Educational Planning.

Jannah, R., Hidayat, D. N., Husna, N., & Khasbani, I. (2021). An item analysis on multiple-choice questions: A case of a junior high school English try-out test in Indonesia. *Leksika: Jurnal Bahasa, Sastra Dan Pengajarannya*, *15*(1), 9-17. https://doi.org/10.30595/lks.v15i1.8768

Karjo, C. H., & Ronaldo, D. (2019). The validity of TOEFL as entry and exit college requirements: Students' perception. In *Proceedings of the Eleventh Conference on Applied Linguistics (CONAPLIN 2018)*, 326–330. https://doi.org/10.2991/conaplin-18.2019.277

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. In *American Journal of Health-System Pharmacy, 65*(23), 2276–2284). https://doi.org/10.2146/ajhp070364

Kunandar, K. (2013). *Penilaian autentik: Penilaian hasil belajar peserta didik Kurikulum 2013*. Raja Grafindo Persada.

Leung, C. (2022). Language proficiency: from description to prescription and back? *Educational Linguistics*, *1*(1), 56–81. https://doi.org/10.1515/eduling-2021-0006

Lia, R. M., Rusilowati, A., & Isnaeni, W. (2020). NGSS-oriented chemistry test instruments: Validity and reliability analysis with the Rasch model. *REID (Research and Evaluation in Education)*, *6*(1), 41-50. https://doi.org/10.21831/reid.v6i1.30112

Maharani, A. V., & Putro, H. N. P. S. (2020). Item analysis of English final semester test. *Indonesian Journal of EFL and Linguistics*, *5*(2), 491–504. https://doi.org/10.21462/ijefl.v5i2.302

Moses, T. (2017). A review of developments and applications in item analysis. In R. Bennett & M. von Davier (Eds.), *Methodology of educational measurement and assessment: The methodological, psychological and policy contribution of ETS*. Springer Open. https://doi.org/10.1007/978-3-319-58689-2

Mouvet, K., & Taverniers, M. (2022). What is language anyway? A view on teaching English proficiency in higher education. *International Journal of TESOL Studies*, *4*(2), 8–23. https://doi.org/10.46451/ijts.2022.02.02

Muchlisin, M., Mardapi, D., & Setiawati, F. A. (2019). An analysis of Javanese language test characteristic using the Rasch model in R program. *REID (Research and Evaluation in Education)*, *5*(1), 61–74. https://doi.org/10.21831/reid.v5i1.23773

Mustafa, F. (2015). Using corpora to design a reliable test instrument for English proficiency assessment. In *The 62nd TEFLIN International Conference 2015*, 344–352. https://repositori.unud.ac.id/protected/storage/upload/repositori/d6117bc1b9d271bd3f1b3fbee69683cc.pdf

Mustafa, F., & Apriadi, H. (2014). DIY: Designing a reading test as reliable as a paper-based TOEFL design by ETS. In *Proceedings of the 1st English Education International Conference (EEIC) in Conjunction with the 2nd Reciprocal Graduate Research Symposium (RGRS) of the Consortium of Asia-Pacific Education Universities (CAPEU)*, 402–407. http://eeic.unsyiah.ac.id/proceedings/index.php/eeic/article/view/79

Ndayizeye, O. (2017). Discrepancies in assessing undergraduates' pragmatics learning. *REID (Research and Evaluation in Education)*, *3*(2), 133-143. https://doi.org/10.21831/reid.v3i2.14487

Ofianto, O. (2018). Analysis of instrument test of historical thinking skills in senior high school history learning with Quest programs. *Indonesian Journal of History Education*, *6*(2), 184–192. https://journal.unnes.ac.id/sju/index.php/ijhe/article/view/27648

Phillips, D. (2001). *Longman introductory course for the TOEFL test*. Longman.

Pratama, D. (2020). Analisis kualitas tes buatan guru melalui pendekatan Item Response Theory (IRT) model Rasch. *Tarbawy : Jurnal Pendidikan Islam*, *7*(1), 61–70. https://doi.org/10.32923/tarbawy.v7i1.1187

Rahim, A., & Haryanto, H. (2021). Implementation of Item Response Theory (IRT) Rasch model in quality analysis of final exam tests in Mathematics. *Journal of Research and Educational Research Evaluation (JERE)*, *10*(2), 57–65. https://doi.org/10.15294/jere.v10i2.51802

Renandya, W. A., Hamied, F. A., & Nurkamto, J. (2018). English language proficiency in Indonesia: Issues and prospects. *Journal of Asia TEFL*, *15*(3), 618–629. https://doi.org/10.18823/asiatefl.2018.15.3.4.618

Rizbudiani, A. D., Jaedun, A., Rahim, A., & Nurrahman, A. (2021). Rasch model item response theory (IRT) to analyze the quality of mathematics final semester exam test on system of linear equations in two variables(SLETV). *Jurnal Pendidikan Matematika*, *12*(2), 399–412. http://ejournal.radenintan.ac.id/index.php/al-jabar/index

Sacko, M., & Haidara, Y. (2018). Developing autonomous listening learning materials for university students TOEFL preparation. *LingTera*, *5*(2), 170–178. https://doi.org/10.21831/lt.v5i2.10192

Saswati, R. (2021). Item analysis of reading comprehension test: A study of test scores interpretation. *Scope : Journal of English Language Teaching*, *6*(1), 42-49. https://doi.org/10.30998/scope.v6i1.7675

Setyawarno, D. (2017). *Panduan penggunaan program Quest untuk analisis butir soal hasil belajar bahasa model konvergen dan divergen*. Universitas Negeri Yogyakarta.

Sugianto, A. (2020). Item analysis of English summative test: EFL teacher-made test. *Indonesian EFL Research and Practices*, *1*(1), 35–54. https://journal.iaima.ac.id/i-efl/article/view/4

Suryani, N. Y., & Khadijah, S. (2021). The effectiveness of virtual classroom in TOEFL preparation. *Acitya: Journal of Teaching & Education*, *3*(2), 198-209. https://doi.org/10.30650/ajte.v3i2.2199

Thu, A. S. (2019). Autonomous learning materials of structure and written expression for TOEFL preparation. *LingTera*, *6*(1), 62–72. https://doi.org/10.21831/lt.v6i1.15919

Thurmond, V. A. (2001). The point of triangulation. *Journal of Nursing Scholarship*, *33*(3), 253–258. https://doi.org/10.1111/j.1547-5069.2001.00253.x

Wahyuni, A., & Kartowagiran, B. (2018). Developing assessment instrument of qirāatul kutub at Islamic boarding school. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *22*(2), 208–218. https://doi.org/10.21831/pep.v22i2.16592

Wright, B. D., & Mok, M. M. C. (2004). An overview of the family of Rasch measurement models. In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement theory, models and applications* (pp. 1–24). JAM Press.

Yumelking, M. (2019). Test items analysis constructed by EFL teachers of private senior high school in Kupang, Indonesia. *International Journal of English Literature and Social Sciences*, *4*(6), 1746–1752. https://doi.org/10.22161/ijels.46.19