International Journal of Multicultural and Multireligious Understanding

http://ijmmu.com
editor@ijmmu.com
ISSN 2364-5369
Volume 8, Issue 12
December, 2021
Pages: 647-654

An Item Analysis of English Test During Online Learning

Umi Ma'rifah[1]; Nyanuar Algiovan[1]; Cucu Sutarsyah[2]

[1] Master of English Education Program, University of Lampung, Indonesia

[2] Professor in English Department, The Faculty of Education and Teacher Training, University of Lampung, Indonesia

http://dx.doi.org/10.18415/ijmmu.v8i12.3396

Abstract

The assessment process in education holds a crucial and fundamental function since the quality of Education will be measured from the assessment process. The process of assessing language teaching then becomes interesting as a study material during online learning. This study aimed to dissect the quality of three important things in the analysis of English test items namely Item Difficulty, Item Discrimination, and Distractors Effectiveness. This study used a quantitative research approach. The research sample was obtained from 49 answer sheet of students' work in taking the school exam. The researchers used Iteman application program to analyze the results. The results revealed that the test used had low reliability. In terms of the item difficulty, the majority of the items (52 %) are difficult, 34% moderate and 14% in easy category. None of the items is considered very good in item discrimination, some of items accepted but revision, and rejected. Whilst, it was found that in the distractors, 38 items are effective and the rest are ineffective. It is concluded that test makers' understanding on the quality of the test are needed to determine the quality of a good and appropriate language measurement.

Keywords: *English Test; Item Analysis; Iteman; Online Learning*

Introduction

Numerous studies with the theme of learning approaches, techniques to English language, and teaching methods have been widely carried out by various researchers, practitioners and experts. In fact, the results of those studies should also be accompanied by a review of assessment pattern of language research trends because assessment is a measuring medium for the application of approaches, techniques and learning methods in the classroom. Assessment becomes important in the learning process (Kearns, 2012), due to the assessment process, eventually the teachers will know the ability, and weaknesses during the learning process (Barbosa & Garcia, 2005; Rahmawanti & Umam, 2019; Maharani & Putro, 2020). It is correlated that a quality test will produce a portrait of the success of the English teaching process (Hamp-Lyons, 2007; Rahmawati & Ertin, 2014; Sato, 2018; Rao & Haque, 2019). In education, assessment involves collecting evidence and making judgments or forming opinions about learners' knowledge, skills, and abilities. It often also involves keeping an informal or formal record of those judgments. It is a key to professional responsibility of all teachers to become critical and effective language skills assessors. Unless teachers are able to judge what learners know or can do, they neither reasonably decide whether their teaching has been successful nor they can choose what to do next to help learners to improve (Inbar-Lourie, 2008).

In Indonesia, assessment in the form of tests is also widely used for several purposes, such as the admission of new students, rate increases, graduation, and grouping of students based on test results. However, some research results found that the quality of the questions are not appropriate with the criteria of a good test (Rahmawati, 2012; Fiktorius, 2014; Jannah et al., 2021).

According to Brown (2001), a test should consist of at least three criteria encompasses validity, reliability, and practicality. Validity refers to the extent that the test is able to measure what should be measured based on the learning goals or competencies to be achieved; while reliability refers to consistency or dependability of the scores obtained. Practicality can broadly be defined as operating budget, time limitation, implementation, and scoring system. Test should be prepared with the low budget (Brown, 2001). Then, test should have vivid time limitation and could be managed easily. The most important to say is the specific and efficient scoring system. Associating with reliability, the test result should provide stable results in different circumstances (Fulcher & Davidson, 2007). Therefore, the test result is trusty. For this reason, item analysis is crucial way to determine the quality of the test items before being used to evaluate the learners' language performance. A good item should conform at least to three characteristics, namely item difficulty, discrimination power, and distractors effectiveness (Brown, 2004).

The first characteristic is item difficulty. This refers to identifying the proportion of students who answer correctly (Haladyna, 2004). This definition is similar to that found in Brown (2004) who writes the item difficulty relates to the percentage of students who assume whether an item is easy or difficult. Second characteristic is item discrimination that has ideal index of more than 0.39 (Ebel & Frisbie, 1991) with range between 0.0 and 1.0 (Hingorjo & Jaleel, 2012). This characteristic is about identifying students' knowledge and ability (Haladyna, 2004). It assists teachers to discover high achievers and low achievers in a class. An item test can reach ideal index when high achievers answer correctly more often than low achievers (Hingorjo & Jaleel, 2012). The third characteristic is distractors effectiveness. This characteristic can only be analyzed on tests in the form of multiple-choice tests. A well distractor must be chosen by at least 5% of the respondents, especially those who include in low achievers (Rosana & Setyawarno, 2017). The distractors effectiveness analysis is one of important parts since it has several functions in item analysis. The functions involve reducing items that use ineffective sentences or too many options, providing information to improve the items, assisting to choose a correct distractor, assisting to comprehend students' cognitive behavior, and increasing items' response score (Haladyna, 2004).

Item analysis, especially in language research has been widely done in Indonesia for Junior High Schools Level (Manfenrius & Sutapa, 2015; Indrayani et al., 2020; Trivict & Densiana, 2020; Jannah et al., 2021; Karim et al., 2021), Senior High Schools (Manalu et al., 2019; Hartati & Yogi, 2019; Maharani & Putro, 2020), and at tertiary level (Triono et al., 2020). However, the study of analysis of details in Vocational Schools has not gained a place in the assessment study on language research. Not to mention, pandemics also change assessment patterns in education (Bashir et al., 2021). Assessments that are commonly done offline switch modes to online, which has an impact on a variety of issues (Yulianto & Mujtahin, 2021). The weak ability of English teachers is also suspected to be a classic problem of test quality (Azis, 2015; Wijayanti, 2019). The researchers then think to fill this gap with the data of vocational high school students' English test with the aims to dissect the quality of three important things in the analysis of English test item namely *Item Difficulty, Item Discrimination, and Distractors Effectiveness.*

Methodology

This study used descriptive quantitative research. The sample involved in this study is 49 samples chosen randomly in the form of students' answer sheets. There are 50 items in the form of multiple-choice test with 4 options in the English final test for Vocational High Schools in East Lampung. The researchers applied a blank table as an instrument in this study. The blank table refers to the *Iteman* program report of multiple-choice test item analysis. The researchers used this table to record the calculation results of *Iteman* program. Since this study aimed to dissect the quality of three important things in the analysis of English test items, namely *Item Difficulty, Item Discrimination and Distractors Effectiveness*, the researchers used *Iteman* program. *Iteman* (Item and Test Analysis) is a Windows application designed to provide a detailed report of the test items and analysis of the test using classical theory.

This instrument involves 3 characteristics: item discrimination, item difficulty, and distractors in which these characteristics cover the quality of the test. To analyze the data, the researchers computed through the *Iteman* to obtain the calculation of item difficulty, item discrimination, and distractors index. The answer key and students' responses of the test package were typed in the form of notepad file. Afterward, the researchers created control file in the form of notepad as a command to analyze the data. The control file must be placed in the same location as the *Iteman* program. Then, the researchers ran the program and typed the 'submit' word followed by the control file's name.

Finding

The Quantitative analysis was run to identify the quality of test items based on item difficulty, item discrimination, and distractors effectiveness using *Iteman* program. Before dissecting the quality of three important things in the analysis of English test item, the researchers explored the descriptive data analysis and reliability of a test first. The description of the research data can be seen in the table below:

Table 1. Descriptive Data Analysis

Score	Items	Mean	SD	Min Score	Max Score	Mean P	Mean Rpbis
Scored Items	50	16.898	3.555	10	29	0.338	0.058
Scaled Total	50	0.000	0.000	0.000	0.000	-	-

The data explicated that the total test items is 50 items, with Mean of 16.898 and a standard deviation is 3.555. The minimum score obtained by students who take the test is 10 while the maximal score is 29 items. The mean of P is 0.338 and mean *Rpbis* is 0.058. The results of *Iteman* attained a distribution of scores which confirmed that only 1 student was able to answer correctly 58% of the questions, while 48 students gained a score below 50% of the total test.

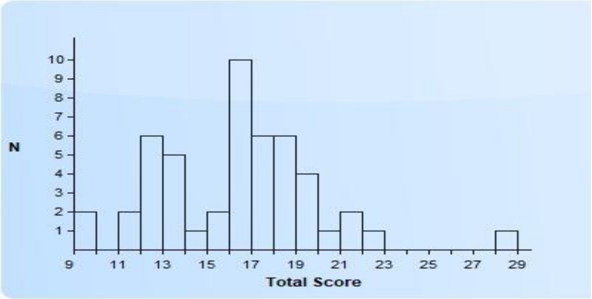

Chart 1. The Distribution of Score

Reliability of the Items

Reliability can be interpreted as a measuring device that has the implication that the test result has relatively the same at different times with the same measuring instrument. In measuring reliability of the items, the researchers used Coefficient reliability criteria adopted by Guilford in the following table.

Table 2. Coefficient Reliability

Index	Category
-1.00 - 0.20	Very Low
0.20 - 0.40	Low
0.40 - 0.60	Moderate
0.60 - 0.80	High
0.80 - 1.00	Very High

Table 3. Reliability of the Data

Score	Alpha	SEM	Split-Half (Random)	Split-Half (First-Last)	Split-Half (Odd-Even)	S-B Random	S-B First-Last	S-B Odd-Even
Scored items	0.317	2.937	0.099	0.316	0.052	0.180	0.480	0.099

Table 3 shows that Alpha is 0.317 with a Standard Error of Measurement of 2.937. Thus, it can be concluded that the data as a whole has low reliability. In addition, Sukardi (2008) argues that the higher Alpha coefficient that the test has, the more consistent the test is. Dealing with this assumption, it can be concluded that reliability of the English Test administered in final examination during an online learning for Vocational High Schools in East Lampung is low. In the other words, the consistency of the test is also low (Sukardi, 2008).

Having known the descriptive data analysis and reliability of a test, the researchers then dissected the results of item difficulty, item discrimination, and distractors effectiveness, as follow:

Item Difficulty

To find out the results of item difficulty, the researchers calculated the student's answer sheets using the *Iteman*. Table 4 displayed the item difficulty.

Table 4. The Item Difficulty of English Test

Index	Category	N	Items Number	Percentage
0.00-0.30	Difficult	26	3,4,5,7,9,10,11,14,17,19,23, 27,28,30,33,34,37,38,40,42, 43,44,47,48,49,50	52%
0.31-0.70	Moderate	17	1,2,8,12,13,15,16,18,20,21 ,22,29,36,39,41,45,46	34%
0.71-1.00	Easy	7	6,24,25,26,31,32,35	14 %
		50		100 %

Iteman also calculated and displayed the item difficulty of the test. It can be concluded that more than half of the items are difficult. The above table confirmed that 26 items or 52% of the test items are in category of the difficult items, 17 items or 34% are in the moderate, and the rest 7 items or 14% are in the easy items.

Item Discrimination

The results analysis of item discrimination was gained using *Iteman* by looking at the index point-biserial correlation coefficient, and then classified it based on the category on the index. The following table shows the item discrimination of English test.

Table 5. The Item Discrimination of English Test

Index	Category	N	Items Number	Percentage
0.40 -1.00	Very Good	0	-	-
0.30-0.39	Accepted but Little revision	6	2,8,13,28,35,50	12%
0.20-0.29	Need revision	8	1,9,25,27,30,31,40,44	16%
< 0.19	Rejected	36	3,4,5,6,7,10,11,12,14,15,16,17,18, 19,20,21,22,23,24,26,29,32,33,34,36 37,38,39,41,42,43,45,46,47,48,50	72%
Total		50		100%

Iteman also calculated and displayed the finding that surprisingly there are no items in index of very good items and the majority of items are rejected. The data above shows that most of the items (36 items or 72%) given to the students are in bad category, meaning that it must be rejected. Meanwhile, only a few items (6 items or 12%) are accepted but requires little revision, and 8 items or 16 % of the total items needed revision to satisfy the standards of items discrimination index.

Distractors Effectiveness

The last aim in this study is to find out distractors effectiveness. To circumscribe the effectiveness of alternative answers as a distractor, the data was gained based on the criteria of at least 5% of test-takers who responded to the alternative answers. Table 6 displayed the results of item analysis on distractor effectiveness for 50 items of English utilized to test the ability of vocational high students in East Lampung.

Table 6. The Effectiveness of Distractors of English Test

Category	Number of Items	Percentage
Effective	38	76%
Ineffective	12	24%
	50	100%

As clearly presents in Table 6, almost all of the distractors of the test had effective category to distract the students. The interesting point of the distractor data from *Iteman* is that 38 items or 76% are effective distractor and the rest 12 items or 24% are ineffective distractor. The results show that most of the items may distract students effectively.

Discussion

Iteman recorded that there are confirmed 52% items as difficult items, 34% items as moderate and 14% items as easy category. The proportion of the test items for vocational high school students in East Lampung can be said not to satisfy the standards of an ideal test. To make it an ideal test, it is necessary to discard 13 items of difficult items category, then add 8 items to the moderate category, and rearrange 5 items in easy category. By doing so, it will be found an ideal test that can be used to see the ability of students. This is in accordance with the statement from Kusnandar as quoted by Amalia and Nur (2020) that an ideal test contains 25% of difficult items, 50% of moderate items and 25% of easy items category, with the ratio of an ideal test is 1:2:1 (Heaton, 1988). Moreover, it is pointed out that the recommendation is frequently made that they include only those test items with midrange level of difficulty between 40 and 70 percent (Ebel & Frisbie, 1991). Thus, the English test used in vocational high schools in East Lampung then still cannot be a parameter in measuring the success and ability of students (Anderson & Arsenault, 1998).

In line with the above explanation, Brown emphasized that the test items which are used to test the ability should cover each difficulty level so that teachers can recognize the abilities of each student. It is also supported by the statement: the test that is used to measure students' achievement should not too easy and not too difficult for students (Arikunto, 2012). The results of this investigation then confirmed the findings of several studies that have been conducted by several researchers (Manalu et al., 2019; Hartati & Yogi, 2019; Trivict & Densiana, 2020; Maharani & Putro, 2020; Jannah et al., 2021; Karim et al., 2021). Factually, many studies reported that the level of difficulty of the test used has been prepared by using the ideal test standards (Indrayani et al., 2020).

Regarding Item difficulty in English test, it is believed that it can be influenced by cognitive factors (Sung et al., 2015). Danili & Reid (2006) argued that cognitive factors involve comprehension, coding, transition, scrutinizing, and working memory. Moreover, they added the cognitive factors that affect students' performance and achievement so these factors affect calculation of item difficulty. Thus, teachers or test makers need to consider cognitive factors in satisfying difficulty index items.

In Item Discrimination, the data reveal that there is no item in a very good category. From the results of item analysis, 36 items or 72% are in bad categories, meaning those must be discarded. Meanwhile, 6 items or 12% are accepted in index discrimination but requires little revisions, and 8 items or 16% require revision to reach the index discrimination of test standards. The results of this study are in line with some previous studies. This phenomenon then becomes a landscape of how the quality of the item of English Test is used to measure students' abilities, especially during online learning.

The test items will be considered having a high discrimination number if the intelligent students succeed in answering the questions correctly and the less intelligent students failed to answer the questions. However, if these two groups are able to answer the same questions correctly, the value of discrimination is zero. Also, if the less intelligent students answer the same questions correctly more than the intelligent one, the value of discrimination is negative. These zero and negative values are the things that should be avoid for the test items.

Meanwhile, the results of item analysis with regard to distractors effectiveness show that distractors in 38 items or 76% are very effective and 12 items or 24% are ineffective or require to be revised. The distractor is supposed to function that if the lower the ability level of test taker, the more distractor would be chosen; or the higher the ability level of test taker, the less distractor would be chosen. This can be shown by the presence of a high, low or negative correlation in the results of the analysis. A test is said to have good discriminating power if it can distinguish between students who have high ability with students who have low ability.

Conclusion

The results of the analysis show that the quality of the test is still in the not ideal category. It can be seen from the data that the test had low reliability, the proportion of difficulty items that did not refer to the standard of an ideal test and the proportion of item discrimination where the majority of items needed to be discarded and revised. However, most of the distractors are effective to use. This study implies a phenomenon of the assessment process that also occurs in any evaluation process. Regarding to this, the researchers advise teachers and test makers to pay attention to every details of the element of test, so that the test has good quality. In addition, studies on the test items analysis need to gain a wide place in language education research, because teachers and test makers' understanding on the quality of the test are needed to determine their ability in measuring students' language skills.

The researchers hope to further studies to investigate the test item analysis on English test using other classic test applications, so that the results of the study can be compared. Moreover, the researchers believe that the sample research used is still small. Testing from a test using broader samples and other research methodology will add to the results of the study of item analysis in further research.

References

Anderson, G., & Arsenault, N. (1998). *Fundamentals of Educational Research*.

Arikunto, S. (1984). Dasar-Dasar Evaluasi Pendidikan, Yogyakarta:PT. BINA AKSARA.

Azis, A. (2015). Conceptions and Practices of Assessment: a Case of Teachers Representing Improvement Conception. *TEFLIN Journal - A Publication on the Teaching and Learning of English*, *26*(2), 129. https://doi.org/10.15639/teflinjournal.v26i2/129-154

Barbosa, H., & Garcia, F. (2005). Importance of online assessment in the E-learning process. *ITHET 2005: 6th International Conference on Information Technology Based Higher Education and Training, 2005*, *2005*(May 2014), 1–7. https://doi.org/10.1109/ITHET.2005.1560287

Bashir, A., Uddin, M. E., Basu, B. L., & Khan, R. (2021). Transitioning to online education in English departments in Bangladesh: Learner perspectives. *Indonesian Journal of Applied Linguistics*, *11*(1), 11–20. https://doi.org/10.17509/ijal.v11i1.34614

Danili, E., & Reid, N. (2006). Cognitive Factors That Can Potentially Affect Pupils' Test Performance. *Chemistry Education Research and Practice*, *7*(2), 64–83. https://doi.org/10.1039/B5RP90016F

Ebel, R. L., & Frisbie, D. A. (1991). Essentials of Educational Measurement Fifth Edition. In *Prentice Hall of India*. https://doi.org/10.1016/0022-4405(73)90057-5

Fiktorius, T. (2014). *A Validation Study on National English Examination of Junior High School in Indonesia*.

Fulcher, G., & Davidson, F. (2007). Language Testing and Assessment: An Advanced Resource Book. In *ELT Journal* (Vol. 63, Issue 2). https://doi.org/10.1093/elt/ccp010

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

Heaton, J. (1988). Writing English Language Tests. In *Longman Group*.

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. In *Language Testing* (Vol. 25, Issue 3). https://doi.org/10.1177/0265532208090158

Indrayani, M. S. D., Marhaeini, A. A. I. N., Paramartha, A. A. G. Y., & Wahyuni, L. G. E. (2020). The Analysis of the Teacher-Made Multiple-Choice Tests Quality for English Subject. *Journal of Education Research and Evaluation*, *4*(3), 272. https://doi.org/10.23887/jere.v4i3.25814

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. https://doi.org/10.30595/lks.v15i1.8768

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

Kearns, L. (2012). Student Assessment in Online Learning: Challenges and Effective Practices. *Jolt.Merlot.Org*, *8*(3), 198–208. http://jolt.merlot.org/vol8no3/kearns_0912.htm

Maharani, A. V., & Putro, N. H. P. S. (2020). Item Analysis of English Final Semester Test. *Indonesian Journal of EFL and Linguistics*, *5*(2), 491. https://doi.org/10.21462/ijefl.v5i2.302

Manalu, D., Sipayung, K. T., & Lestari, F. D. (2019). an Analysis of Students Reading Final Examination By Using Item Analysis Program on Eleventh Grade of Sma Negeri 8 Medan. *JETAL: Journal of English Teaching & Applied Linguistic*, *1*(1), 13–19. https://doi.org/10.36655/jetal.v1i1.98

Manfenrius, A., & Sutapa, G. B. W. (2015). *Items Analysis on the Score of the English Summative Test*. 1–10.

Rahmawanti, M. R., & Umam, A. (2019). Integrating Web 2.0 Tools in Writing Class to Promote Assessment for Learning. *JEES (Journal of English Educators Society)*, *4*(2), 53. https://doi.org/10.21070/jees.v4i2.2516

Rahmawati, Y., & Ertin. (2014). Developing Assessment For Speaking. *Indonesian Journal of English Education*, *1*, 1–12.

Rahmawati, Y. (2012). *English education department teacher training and education faculty sebelas maret university surakarta*. *265*, 265–278.

Rao, M. M., & Haque, M. I. (2019). A study on impact of testing on English as a foreign language teaching in a Saudi Arabian University. *Humanities and Social Sciences Reviews*, *7*(2), 58–71. https://doi.org/10.18510/hssr.2019.727

Sato. (2018). The Impact of the Test of English for Academic Purposes (TEAP) on Japanese Students' English Learning. *JACET Journal*, *62*, 89–107.

Sung, P.-J., Lin, S.-W., & Hung, P.-H. (2015). Factors Affecting Item Difficulty in English Listening Comprehension Tests. *Universal Journal of Educational Research*, *3*(7), 451–459. https://doi.org/10.13189/ujer.2015.030704

Triono, D., Sarno, R.., & Sungkono, K. R. (2020). Item Analysis for examination test in the postgraduate student's selection with classical test theory and rasch measurement model. *Proceedings - 2020 International Seminar on Application for Technology of Information and Communication: IT Challenges for Sustainability, Scalability, and Security in the Age of Digital Disruption, ISemantic 2020*, 523–529. https://doi.org/10.1109/iSemantic50169.2020.9234204

Trivict, T., & Densiana, F. (2020). *The quality of an English summative test of a public junior*. *3*(2), 133–141.

Wijayanti, D. N. (2019). English Teachers' Understanding of Language Assessment. *Journal of English Teaching and Learning Issues*, *2*(2), 93–114. https://doi.org/10.21043/jetli.v2i1.

Yulianto, D. & Mujtahin, N. M. (2021). Online Assessment during Covid-19 Pandemic : EFL Teachers ' Perspectives and Their Practices. *Journal of English Teaching, 7(2), 229-242. DOI: Https://Doi.Org/10.33541/Jet.V7i2.2770*, *7*(2021), 229–242. https://doi.org/10.33541/jet.v7i2.2770

Copyrights