World Journal on Educational Technology: Current Issues

Volume 14, Issue 4, (2022) 1025-1037

Evaluation of an online teacher-made test through blackboard in an English as a foreign language writing context

Mohd Nazim [*a], Najran University, Department of English, College of Language & Translation, Saudi Arabia

Ali Abbas Falah Alzubi [b], Najran University, Department of English, College of Languages & Translation, Saudi Arabia

Suggested Citation:

Nazim, M. & Alzubi, A. A. F. (2022). Evaluation of an online teacher-made test through blackboard in an English as a foreign language writing context. *World Journal on Educational Technology: Current Issues*. *14*(4), 1025-1037. https://doi.org/10.18844/wjet.v14i4.7614

Received from March 13, 2022; revised from May 16, 2022; accepted from July 25, 2022.
Selection and peer-review under responsibility of Prof. Dr. Servet Bayram, Yeditepe University, Turkey.

Abstract

The significance of this study is to improve the assessment quality of future online English as a Foreign Language (EFL) writing tests through Blackboard, a learning management system (LMS), and to avert any potential inclusion of odd items. This study aims to examine the Blackboard (Bb) test quality at the Preparatory Year Program (PYP) in an English for Specific Purposes (ESP) writing course using item analysis and a questionnaire of EFL teachers' practices for constructing a good quality test. To achieve the study objectives, 30 objective-type questions from the final Technical Writing course examination, attempted by 97 level two preparatory year students, were analyzed to check three indices: difficulty index, discrimination index, and distractor efficiency. In addition, a questionnaire was administered to rate the EFL teachers' (N=50) practices of constructing their technical writing test for the final examination in terms of good quality test norms. The item analysis has shown that the test proved to be valid and reliable; however, it was easy. Many items had no discrimination indices, and many distractors were not functioning. The analysis of the questionnaire data showed that there was a high level of commitment by the EFL teachers to apply the required norms for constructing a good quality language test. Genders and teaching experience had no significant differences in the EFL teachers' degree of the test norms employment. In the light of the findings, recommendations and further research are suggested.

Keywords: Blackboard, evaluation, good quality test, Saudi EFL writing context, teacher-made test

* ADDRESS OF CORRESPONCE: Mohd Nazim, Najran University, Department of English, College of Language & Translation, Saudi Arabia
Email address: nazimspeaking@yahoo.co.in

1. Introduction

One of the essential phases in language learning is the accurate assessment of students' learning outcomes. This requires that the assessment tools must be valid and reliable. Testing, one of the most applied measurement tools, is very essential in the development of students' communicative competence (Basanta, 2012; Saragih, 2015). In fact, they are necessary for assisting English as a Foreign Language (EFL) teachers to collect evidence from the test-takers, so as to make inferences about the test takers' language knowledge or language usability (Fulcher & Davidson, 2007). Due to the sudden spread of COVID-19, the emergent transform to be completely online education using learning platforms such as Blackboard has posed various challenges that have required urgent solutions, such as the online education efficiency, motivation, integration, use of technology, and assessment. The challenge of the efficiency of similar offline tests to measure the students' really performance levels arises amid the conduction of the test remotely and online. That means students take the test without any direct surveillance that may raise the concerns of teachers on their progress (Fitriyah & Jannah, 2021; Ghanbari & Nowroozi, 2021).

1.1. *Theoretical and conceptual framework*

There is a lack of evaluative research on item analysis of teacher-made tests in EFL writing contexts in online learning environments. Therefore, the study is of significance to evaluate a teacher-made test in the Saudi EFL writing context in terms of discrimination, difficulty, and efficiency of distractors. The study also rates the EFL teachers' practices of constructing a good language test.

1.2. *Related research*

The educational process consists of many tasks of which teaching and assessment may be the most important. Basanta (2012, p. 40) states that "teaching and testing are two inseparable units of the teacher's task". He asserts that teachers should assess students based on what is taught to them, and there should be a sort of harmony between the test usage and the students' ability to use language appropriately.

Students' language assessment is a sensitive task that requires matching between some input items such as instructional objectives, learning outcomes, and study program. Also, validity is considered the main concept in language assessment (Fulcher & Davidson, 2007). A language test being one of the assessment tools used in language education should be valid and reliable. It should also be able to discriminate between students' levels. The test is defined as a tool that includes questions aiming to measure students' behavior such as level of achievement in certain skills after a particular period (Allen & Yen, 1979; Brown, 2003; Gronlund & Linn, 2000; Lebagi et al., 2017). EFL teachers always measure students' progress and knowledge through a test, the most widely used assessment tool. By doing so, teachers should be familiar with the norms for constructing a good language test (Aain et al., 2020). Besides, a teacher-written test can be reexamined in terms of being well-written through analyzing the students' responses to the test's items. This stage includes checking the test reliability, level of difficulty, discrimination index, and efficiency of distractors. In this regard, item analysis is a relevant exercise to highlight strong and weak points of the test through statistics, and thus to improve in a later version of the test (Ahman & Glock, 1971; McCowan & McCowan, 1999).

Students' writing requires assessment in order to have evidence on their level of advancement (Dolin & Evan, 2018). The process of EFL writing assessment includes eight areas: content, organization, discourse, syntax, vocabulary, and mechanisms (Frazier & Brown, 2001).

These areas fall under various types of EFL writing assessment including clear explanation, feedback, rubrics, interaction with teachers and peers (Gaytan & McEwen, 2007). These forms of EFL writing assessment can be used to assess students' writing abilities in face-to-face and online learning environments. For example, the use of assessment in form of feedback enhanced students' writing skills (Fatimah & Yusuf, 2019). However, the application of such forms online may be a problem for many teachers. Dwiyanti and Suwastini (2021) claimed that although teachers' understanding of the significance of EFL writing assessment, many areas related to writing organization, discourse, and mechanisms were not assessed due to their lack of experience in using e-learning platforms.

In the teaching and learning process, it is expected that learners who have completed a unit of learning tasks should be different in terms of knowledge acquisition or course learning outcomes from those who have not done it. It is useful for a teacher to conduct a test to assess their competence or learning outcomes. In general, the teacher as the test maker pays attention to writing a good quality test following a set of criteria. Research on validating and making online EFL teacher-made tests reliable through item analysis is very limited. The problems of testing have ranged from the relatedness of the test to the study program (Putri, 2009), to students' mere focus on tests rather than language learning process (Lebagi et al., 2017), to issues such as discrimination and distraction efficiency (Toksöz & Ertunç, 2017), and some violations in the norms of a good quality English test (Aain et al., 2020).

Putri (2009), who reported on evaluating 50 school students' answers on a summative multiple-choice English subject test developed by a teacher in Indonesia, found that the questions were related to the curriculum, but did not connect to the students' study program. Also, the test reported no validity and required some revisions. 16% of items were categorized into difficult items, 50% of items belonged to the moderate category, and 34% of items belong to easy items. In addition, it was found that the coefficient of reliability of the test items was 0.841. Also, Lebagi et al. (2017) evaluated an EFL teacher-made test and 10 students' answer sheets at an Indonesian school through an interview, observation, and document analysis as the techniques of collecting data. The results revealed that the 40-item test was reliable (0.7). 92.5% of items had discrimination indices, 25% of distractors did work well. It included 5% difficult items, 27.5% moderate items, and 67.5% easy items. Also, the test influenced the students' motivation to work harder to improve their language ability; however, it may force them to stop learning language and focus only on their test. In addition, Toksöz and Ertunç (2017) reported on analyzing 453 answer sheets of an admission test by English departments using such tools as item facility, item discrimination, and distractor efficiency. The findings found that the multiple-choice test had acceptable item facility indices. However, some items were not discriminative, and some item distractors did not work efficiently. Finally, Aain et al. (2020) assessed the quality of a teacher-developed test using multiple choice in the English subject via a checklist and an interview at an Indonesian school. The result showed the test had 79 very good items and one item as good. However, some items related to punctuation and capitalization violated the norms of a good quality English test.

1.3. *Purpose of the study*

In relation to the previous research, this study is unique in addressing the issue of students' assessment online in EFL context amid the growing need for online education due to the pandemic of COVID-19. Therefore, the study is designed to evaluate teacher-made tests through Blackboard (Bb) using item analysis at the Preparatory Year Program (PYP) in the Saudi EFL writing context.

Also, it examines the EFL teachers' practices for constructing a good quality language test. The following research questions guided this inquiry:

1. To what extent is the teacher-made Bb EFL Technical Writing test valid and reliable?

2. Does the test discriminate students based on how well they know the contents being tested?

3. Do EFL teachers apply the norms of constructing a good quality language test?

4. Is there a significant difference between EFL teachers' practices of writing a good quality test and genders and teaching experience?

2. Method and Materials

This section presents research model, participants, data collection tools, data collection process, and data analysis.

2.1. *Research design*

The study followed the descriptive design for its relevance to the nature of the current study. The study aimed to assess teacher-made tests through Blackboard using item analysis at the Preparatory Year Program in the Saudi EFL writing context. Also, it examined the EFL teachers' practices for constructing a good quality language test.

2.2. *Participants*

The study was conducted at the preparatory year, Najran University. The study had two samples: students and teachers. All level two students were involved in random sampling. As a result, 97 out of 154 male students were selected and participated in the study. The participants were level two preparatory year students who aspire to join scientific majors such as computer sciences, engineering sciences, and medical sciences. At the Preparatory Year, enrolled high school graduates in the scientific stream spend over one year to study a number of courses such as English skills, computer skills, mathematics, and communication skills.

Also, a sample of 50 out of 67 EFL teachers (male=31, female=19) responded to a questionnaire on their practices for writing a good language test. Teachers were selected on a voluntary basis, clearly stated in the questionnaire. And the teachers' approval was achieved through their completing and returning of the questionnaire. The teacher participants are EFL teachers working in the Preparatory Year Program at Najran University. They teach the four English skills to the EFL students over two semesters. That is the total period of the program aiming at qualifying students for specialized majors as illustrated above. They also design the assessment means to measure the students' progress and performance based on a set of criteria on building good quality tests. The teachers hold master and Ph.D. degrees in linguistics, applied linguistics, translation, teaching English as a second language, and English language teaching. Their ages range between (30-65). They have different nationalities: Pakistan, Jordan, Yamen, Egypt, Sudan, India, and Saudi Arabia.

2.3. *Data collection tools*

2.3.1. Test

The student participants attempted to answer 30 objective-type questions from the final Technical Writing (151 TW-2) exam within the allotted time of two hours. The questions had four categories: 12 multiple-choice questions (MCQs), six yes/no questions (YNQs), six multiple answers (MAs), and six matching questions (MQs) as detailed below:

- MCQs: four options (the correct answer and three distractors),

- YNQs: only two options (the correct answer and a distractor),

- MAs: five options (two correct answers and three distractors),

- MQs: nine items and four additional distractors.

Each correct answer is awarded one mark and no penalty marks for wrong answers.

The test was piloted to an exploratory sample of 20 students. The test validity was checked through internal consistency using Pearson correlation analysis. The analysis of the Pearson correlation between each item and the test overall was significant at 0.01 and 0.05. Also, reliability was measured using the Kuder-Richardson formula. It was revealed that the test scored reliability of 0.85, very good and appropriate.

2.3.2. A questionnaire

A questionnaire of good quality test norms was used to examine the teachers' actual procedures for a good language test. An adapted questionnaire including necessary norms for constructing a good quality test (Haladyna, 2004) was used to assess the teachers' commitment to writing a good quality language test through responding to the items on a three-Likert-scale (rarely, sometimes, always). The questionnaire consisted of two main sections: background information and the main items. Background information included the aim of the questionnaire and directional information, an approval statement to participate in the study, the respondent's name, gender (male, female), and teaching experience (-10 years, +10 years). The main items consisted of 10 items of norms for constructing a good quality language test. The norms are related to the form and content of a good language test. The means of the responses were classified within the following range: low (1-1.66), medium (1.67-2.33), and high (2.34-3).

The questionnaire was first checked for reliability using the Kuder-Richardson formula. The analysis showed that the questionnaire had a reliability of 0.82, a very good and appropriate.

2.4. *Data collection process*

Procedures for data collection and analysis

An approval letter for the study conduction was obtained from the Deanship of Scientific Research, Najran University. The approval included the deployment of the questionnaire link, using Google Forms among the targeted teachers, getting access to the students' test data, and analyzing them.

2.4. *Data analysis*

The participants' responses to each item of a summative test were analyzed using SPSS program version 25.1. The test validity was checked through internal consistency using Pearson correlation analysis. Also, reliability was measured using the Kuder-Richardson formula. According to Runder and Schafer (2002), consistency should score 0.50 or 0.60 to demonstrate an acceptable rate of reliability. Also, the questions' level of difficulty was measured using p-value (the number of correct answers (R) on the number of total responses (T'). Discrimination index (DI) was calculated using the upper group of students who received the highest marks subtracted from the lower group of students who received the lowest marks divided on the total number of the two groups. The efficiency of distractors was measured for the test quality. The discrimination indices were compared with the guidelines proposed by Ebel (1979). The importance of DI lies in its power to discriminate between the students' levels. That is to say the higher the DI is, the better the

item is. The distractor efficiency (DE) was calculated based on functioning (chosen at least by 5% of students) and non-functioning items (not chosen by any student). DE is of significance in the quality of items and test results.

3. Results

Research question 1*: To what extent is the teacher-made Blackboard EFL Technical Writing test valid and reliable?*

The test validity was checked through internal consistency using Pearson correlation analysis. Also, reliability was measured using the Kuder-Richardson formula. According to Runder and Schafer (2002), consistency should score 0.50 or 0.60 in order to demonstrate an acceptable rate of reliability. Pearson correlation coefficients were computed between every test question and the test overall for all the students' 97 answers. Table 1 shows the analysis of the Pearson correlation between each item and the test overall at the significance levels of 0.01 and 0.05.

Table 1. Test validity

Item	Pearson correlation	Item	Pearson correlation	Item	Pearson correlation
1	.402**	11	.640**	21	.284**
2	.488**	12	.370**	22	.396**
3	.496**	13	.313**	23	.256*
4	.491**	14	.470**	24	.492**
5	.547**	15	.477**	25	.470**
6	.524**	16	.362**	26	.334**
7	.479**	17	.538**	27	.657**
8	.461**	18	.386**	28	.382**
9	.506**	19	.398**	29	.484**
10	.487**	20	.453**	23	.682**

***. Correlation is significant at the 0.01 level (2-tailed).*

**. Correlation is significant at the 0.05 level (2-tailed).*

As evident in Table 1, all the test items are significant either at the level (0.01) or (0.05), and this indicates that the test is internally consistent. Also, the test was checked for reliability using the Kuder-Richardson formula, and it was revealed that the test scored reliability of 0.85, high and appropriate.

Research question 2: *Does the test discriminate students based on how well they know the contents being tested?*

The psychometric properties of the final EFL Technical Writing test items were computed to check for efficiency and quality. The test (N=30 items) was administrated to 97 students who were split into two classes: high and low according to the overall grade of the test. The answers of 27% of high achievers and 27% of low achievers totaling 26 students were tested. According to the classical theory in evaluation and assessment, the difficulty index is calculated based on the

number of students who answer the item correctly. The value of the difficulty coefficient ranges between 0-1; the more difficult the value is, the easier the item is and vice versa (Nieminen et al., 2010).

The item discrimination and difficulty indices were measured. The analysis showed that the overall difficulty value scored 0.86, indicating that the test was very easy. The difficulty values of the items ranged between 0.77–1.00 (Audah, 2005).

The discrimination index refers to the ability of the test to discriminate between the students who have the knowledge to answer correctly. The discrimination index was measured using the upper group of students who received the highest marks subtracted from the lower group of students who received the lowest marks divided on the total number of the two groups. Any item that received a value of 0.20 and above was considered to have had accepted discrimination, whereas the items receiving less than 0.20 had low discrimination. The following Figure 1 depicts the results of the test analysis.

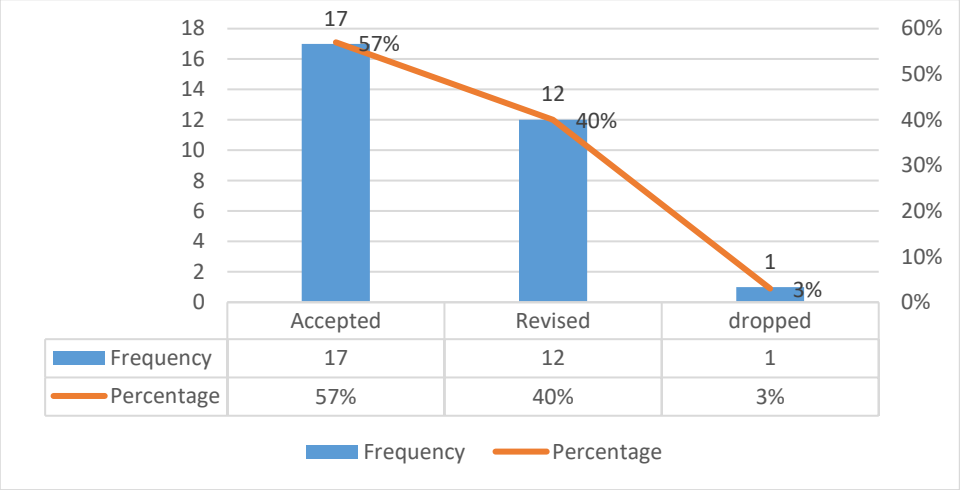

Figure 1. Discrimination index analysis

As shown in Figure 1, it is evident that the discrimination index values scored between 0.00-0.46. Twelve items (40%) received low values of discrimination. Seventeen items (57%) received an acceptable level of discrimination. Only one item (3%) received no discrimination. The zero and low values of the discrimination index may be attributed to a number of factors such as the new environment of test conduction through Blackboard that may have allowed for chances of cheating as students sat for the exam remotely without any surveillance. Also, the sudden change in the mode of education and EFL teachers' less experience of building a good quality test lessened the chances to discriminate between the students' knowledge levels.

The distractor index refers to the wrong options of an item, and for a distractor to be functioning, at least 5% of the students must choose it. It was calculated based on the number of students in the high class who chose the distractor subtracted from the number of students in the low class who choose the distractor divided by the total number of students in both classes. The data analysis of the students' answers revealed that 38 (42%) distractors were functioning. However, 52 (58%) distractors were non-functioning.

The less efficiency of difficulty and discrimination indices may be attributed to some reasons such as the remote application of the test via Blackboard and type of test questions (objective types) that would have given students a chance for cheating and impersonation and teachers' less experience in utilizing the options of building tests through the question system available on Blackboard.

Research question 3: *Do EFL teachers apply the norms of constructing a good quality language test?*

Fifty EFL teachers' perceptions of writing a good quality EFL test were investigated through a questionnaire. The descriptive analysis was applied to extract the EFL teachers' responses to the questionnaire items such as means, standard deviations, and level as shown in Table 2.

Table 2. Descriptive statistics of EFL teachers' practices of writing a good language test

	Item	N	M	SD	level
	Do you ensure that the test follows the general and specific attributes required for preparing language tests?	50	2.80	.404	High
	Do you ensure that the test measures the achievement of the desired objectives and learning outcomes?	50	2.74	.527	High
	Do you ensure that the test measures what it is supposed to measure?	50	2.74	.487	High
	Do you ensure that the test covers all items in the syllabus and integrates other course planning activities?	50	2.66	.557	High
	Do you ensure that the test includes communicative language and is free from nonfunctional material and ambiguous terms?	50	2.68	.471	High
	Do you ensure that the test is neither too difficult (i.e. items are hardly known to students) nor too easy (i.e. one item doesn't aid in answering another) but progressive in difficulty?	50	2.70	.505	High
	Do you ensure that the test is appropriate in length and compatible with the allocated time?	50	2.80	.404	High
	Do you ensure that the test instructs students (for all sections) what to do exactly in a clear way?	50	2.82	.438	High
	Do you ensure that the test is easy to be conducted and scored without wasting too much time or effort?	50	2.78	.418	High
0	Do you ensure that the test produces the same results if given twice to the same students under the same conditions?	50	2.46	.646	High
	Total	50	2.72	.311	High

Table 2 indicates that there is a high degree (M=2.72, SD=.311) in the norms of constructing a good quality language test amongst EFL teachers. That is to say, EFL teachers do consider the

qualities of a good test, match them with the already built language test, and correct any violating norms if there are any. Instructing students about what to do exactly in a clear way had the highest score (M=2.82, SD=.438). Also, the items on ensuring that the test follows the general and specific attributes required for preparing language tests, and the test is appropriate in length and compatible with the allocated time received high equal means and standard deviations (M=2.80, SD=.404). However, ensuring the test production of the same results if taken twice by the same students under the same conditions received the lowest score (M =2.46, SD =.646). In achievement tests, it may be difficult to ensure the validity of the test as students are exposed to the test, and thus there will be a chance for memorizing some items and avoiding mistakes the second time. This action has been in the assignments tests through Blackboard where students were given two attempts. In the second attempt, the results improved better than in the first round.

Research question 4: *Is there a significant difference between EFL teachers' practices of writing a good quality test and genders and teaching experience?*

Kolmogorov-Smirnov Test was used to check the normal distribution of data in relation to the variables of gender and teaching experience. The analysis showed that the data were distributed normally. Therefore, parametric ANOVA analysis was used to check for any significant differences between the 50 (male=31, female=19) EFL teachers' means of norms of writing a good language test attributed to genders and teaching experience (-10 years = 27, +10 years = 23) as displayed in Table 3.

Table 3. EFL teachers' norms of writing a good language test attributed to genders and teaching experience

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	344.996	1	344.996	3601.752	.000
Gender	.165	1	.165	1.719	.196
Teaching Experience	.026	1	.026	.270	.606
Error	4.502	47	.096		
Total	374.110	50			

Table 3 shows that there are no significant differences at the level (0.05) of the EFL teachers' practices of writing a good quality language test attributed to the variables of genders and teaching experience. This is mostly attributed to the good training that the EFL teachers received on constructing a good test.

4. Discussion

Although the writing test proved to be valid and reliable, the overall difficulty indicated that it was very easy. In other words, this test failed to discriminate between students' performance levels. This is a high violation of the test-building criteria and goals. Furthermore, the distractor efficiency varied in function. 42% of the distractors functioned while 58% did not non-function. The findings reported in this study on the easiness of the test and inefficiency of distractors and discrimination indices in many items are in line with the findings by Paramartha's (2017) that

revealed that the EFL reading test taken by a sample of Indonesian school students was easy, and less than half of the test items had eligibility.

The findings also revealed that the EFL teachers had high perceptions of the required norms for constructing a good quality language test during and after the test preparation. Their most distinctive feature focused on ensuring that the test clearly instructs students what to do exactly. It was also found that the EFL teachers' practices of writing a good quality language test did not differ in relation to the variables of genders and teaching experience. This can be linked to the good training that the EFL teachers received on constructing a good test that meets the required standards. In general, this finding does not accord with that by Kabil and Abduh's (2017) study that showed the assessment standards were applied moderately by faculty members in the same context, Najran University. Also, the current study results regarding the differences attributed to genders are inconsistent with those of the study by Kabil and Abduh's (2017) which reported significant differences according to genders in favor of males in the employment of assessment standards.

Based on the triangulation of the results revealed from the students' test analysis and EFL teachers' norms of constructing a good quality language test, it can be noticed that EFL teachers adhere to the required standards for writing a good language test. And this adherence was reflected on the levels of the test validity and reliability. However, the test results showed that the test failed to discriminate between the students' progress and performance levels. This contrast can be attributed to the conduction of the test that was online and remote via the Blackboard. This mode could have allowed some students to seek other means of answering the test such as cheating and impersonation.

Therefore, reasons for a test to be considered as evident from the results as 'easy' on the one hand could be attributed to a number of points. The distractors of the questions were not of that quality which makes them qualify as a distractor. The distractors of the questions were in the role of clue contrary to their quality, i.e., "distracting" students with partial knowledge due to the plausibility to yield the correct option. Also, the nature of objective questions might played a role in the easiness of the exam. Multiple-choice questions may not fit alone in Technical Writing course as students are required to practice various genres of letter, resume, and report, to paraphrase, to make notes, and write a paragraph. However, the teachers' less experience to mark subjective questions following a rubric on the Blackboard platform has limited the assessment of students using essay-type questions in mid-term and final exams. In addition, the similar or same questions were, perhaps, already practiced in the classroom or exercised as home assignments. The questions were designed as teaching points rather than assessment points. Furthermore, the students perhaps are nowadays habitual to pass Blackboard exams with high scores especially for the last two years as compared to the recent past when they used to appear in-person exams. Finally, the students, as news circulates (and chances are higher for sure), take help from others either for the correct answers or impersonation.

However, students' efforts in doing exams should not be ignored. Students who regularly prepare for exams probably will score high. Also, students who are exposed to the exam pattern and variety of questions in the class will see the exam easier and more thoroughly. In addition, students who used to pass exams frequently no matter what forms of assessment they are involved in either formative or no grade, as compared to those who do not appear for an exam at all, will have the results with high scores. Moreover, students who consider themselves lucky are likely to say they scored high because the easy questions came up or the time was favorable to score high. Furthermore, students had no problems like fatigue, stress, or illness on the day of the

exam in addition to their excellent preparation. Finally, students had access to practice to make the study material simple to score high in the test.

In the light of the current findings of this study and the review of literature, some suggestions for better environments of EFL tests employment online are presented. The time allotted for tests online should be carefully calculated to avoid giving students a chance to cheat. Also, essay-type questions would be a choice to reduce the cheating probability in online exams (Ghanbari & Nowroozi, 2021). In addition, the issue of cheating concerns in online exams can be relatively secured by conducting the exams inside the university campus using the available sources of the Internet, computer labs, and networks. Finally, online exams should be skill-focused approach rather than knowledge-based approach. Knowledge has become available at hands. Students need to learn how to reach the knowledge.

5. Conclusion

Teaching and evaluation are inseparable as the effectiveness of teaching is measured through evidence from students being evaluated. This article has evaluated an online teacher-made EFL final Technical Writing test through Blackboard, an LMS, in terms of the efficiency of three indices: difficulty, discrimination, and distraction. The item analysis has shown that the test was valid and reliable, however, it was easy, many items had no discrimination indices, and many distractors were not functioning. This finding may raise doubts on the ability of the test to actually assess the students' language learning knowledge and recommends improvements. It was found that there was a high level of commitment by the teachers to apply those norms in the process of constructing the test. Genders and teaching experience had no significant differences in the EFL teachers' degree of the test norms employment. The study is limited by the number of participants, which may decrease the chances of findings generalization. Also, the study is confined to the EFL Technical Writing course.

6. Recommendations

This research implicates the enhancement of the overall quality of an EFL electronic test via Blackboard through shedding the light on the qualities of a good language test and problems, and suggestions for improvement. It is recommended that:

- EFL teachers analyze the test to evaluate its quality in terms of difficulty index, discrimination index, and distractor efficiency to ensure that the test can cover and measure all the abilities required from students in the textbook as preparatory year is decisive for students who compete for their science majors.
- there is a dire need to train teachers on the various ways of online EFL writing assessment.
- further research on analyzing the test items using a checklist is encouraged to ensure the EFL teachers' employment of the required norms for constructing a good language test.

In addition, evaluating the quality of a test in traditional and online learning modes is suggested.

Acknowledgment: The authors are thankful to the Deanship of Scientific Research at Najran University for funding this work under the General Research Funding program grant code (NU/-/SEHRC/10/1152).

References

Aain, M., Aagy, P., & LGE, W. (2020). The Analysis of the Teacher-Made Multiple Choice Tests Quality for English Subject. *Journal of Education Research and Evaluation, 4*(3), 272 – 278. https://doi.org/10.23887/jere.v4i3.25814

Ahman, J.S. & Glock, M.D. (1971). *Evaluation Pupil Growth, Principle of Tests and Measurements*, 4th Ed. Allyn and Bacon. Inc.

Allen, M.S. & Yen. (1979). *Introduction to Measurement Theory*. Monterey: brooks/Cole.

Audah, A. (2005). *Measurement and Evaluation in the Teaching Process*. Dar Alfikr Press.

Basanta, C. P. (2012). Coming to Grips with Progress Testing: Some Guidelines for Its Design. *English Teaching Forum* 50 (3), 37–40.

Brown, H.D. (2003). *Language Assessment: Principles and Classroom Practices*. Longman.

Dolin, J., & Evans, R. (Eds.). (2018). *Transforming Assessment. Contributions from Science Education Research*. https://doi.org/10.1007/978-3-319-63248-3

Dwiyanti, K. E., & Suwastini, N. K. A. (2021). Assessment for Writing Skills in Online Learning. *Lingua Scientia, 28*(1), 8–19.

Ebel, R. L. (1979). *Essentials of Educational Measurement* (3rd ed.). Prentice Hall.

Fatimah, F., & Yusuf, F. N. (2019, June). Assessment for Learning Impacts on Students' Responsive Writing Skills. *In Eleventh Conference on Applied Linguistics (CONAPLIN 2018)* (pp. 430–435). Atlantis Press. https://doi.org/10.2991/conaplin-18.2019.83

Fitriyah, I., & Jannah, M. (2021). Online Assessment Effect in EFL Classroom: An Investigation on Students and Teachers' Perceptions. *IJELTAL (Indonesian Journal of English Language Teaching and Applied Linguistics)*, *5*(2), 265–284. https://doi.org/10.21093/ijeltal.v4i2.473

Frazier, S., & Brown, H. D. (2001). Teaching by Principles: An Interactive Approach to Language Pedagogy. *TESOL Quarterly, 35*(2), 341–365. https://doi.org/10.2307/3587655

Fulcher, G., & Davidson, F. (2007). *Language Testing and Assessment*. Routledge.

Gaytan, J., & McEwen, B. C. (2007). Effective Online Instructional and Assessment Strategies. *The American Journal of Distance Education, 27*(3), 117–132. https://doi.org/10.1080/08923640701341653

Ghanbari, N., & Nowroozi, S. (2021). The practice of online assessment in an EFL context amidst COVID-19 pandemic: views from teachers. *Language Testing in Asia*, *11*(1), 1–18. https://doi.org/10.1186/s40468-021-00143-4

Gronlund, N. E. & Linn, R. L. (2000). *Measurement and Assessment in Teaching*, 8th Ed. Prentice-Hall, Inc.

Haladyna, T. M. (2004). *Developing and Validating Multiple-Choice Test Items*. Lawrence Erlbaum Associates.

Kabil, R., & Abduh, Y. B. (2017). The Degree of Employment of Faculty Members for Assessment Standards Defined by the American Educational Organizations in Assessing Student Learning at the University of Najran. *Universal Journal of Educational Research, 5*(3), 408–419. https://doi.org/10.13189/ujer.2017.050313

Lebagi, D., Sumardi, & Sudjoko. (2017). The Quality of Teacher-Made Test in EFL Classroom at the Elementary School and Its Washback in the Learning. *Journal of English Education, 2*(2), 97 – 104. https://doi.org/10.31327/jee.v2i2.289

McCowan, R. J., & McCowan, S.C. (1999). *Item Analysis for Criterion-Referenced Test*. SUNY.

Nieminen, P., Savinainen, A., & Viiri, J. (2010). Force Concept Inventory-based multiple-choice test for investigating students' representational consistency. *Physical Review Special Topics-Physics Education Research, 6*(2), 020109. https://doi.org/10.1103/physrevstper.6.020109

Paramartha, A. G. Y. (2017). The Analysis of Multiple-Choice Test Quality for Reading III Class in English Education Department, Universitas Pendidikan Ganesha Bali, Indonesia. *Journal of Education Research and Evaluation, 1*(1), 46–56. https://doi.org/10.23887/jere.v1i1.10060

Putri, Y. F. D. R. (2009). *Analysis of Teacher-Made English Final Second Semester Test for the Year Eleven Students of Sman 1 Ambarawa in the Academic Year of 2008/2009 Based on the Representativeness of Content Standard* (Doctoral dissertation, Universitas Negeri Semarang).

Rudner, L. M., & Schafer, W. D. (2002). *What Teachers Need to Know about Assessment*. National Education Association.

Saragih, F. H. (2015). Testing and Assessment in English Language Instruction. *Jurnal Bahas Unimed, 27*(1), 74–79.

Toksöz, S., & Ertunç, A. (2017). Item Analysis of a Multiple-Choice Choice Exam. *Advances in Language and Literary Studies, 8*(6), 141–146. https://doi.org/10.7575/aiac.alls.v.8n.6p.141