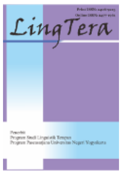

Available online at: http://journal.uny.ac.id/index.php/ljtp

LingTera, 11 (2), 2024, 145–158

Item analysis of pre-service EFL teachers' formative test in teaching practicum for program evaluation

Neti Hartati
Universitas Muhammadiyah Prof. Dr. Hamka, Indonesia
Email: netimpd88@gmail.com

Abstract

Teaching practicum is crucial for pre-service EFL teachers' professional development. Many studies have investigated student-teachers' performance in teaching practicum in various aspects. However, research on their performance in constructing test items in teaching practicum has not been found. Test construction is an essential pedagogical competence that student-teachers must master. Investigating their performance in test construction may give valuable feedback to student-teachers and for teacher education. To fill the research gap, this study conducted an item analysis of a formative test constructed by two student-teachers in a teaching practicum. It revealed that the DP of the items have good quality but the FV analysis showed there is no difficult item suggesting that they cannot make difficult questions. The test reliability is high (0.900625) but there are 6 invalid items. Three items contain grammatical errors creating students' confusion influencing the FV, DP, and the validity. It revealed that the test makers' grammar mastery may determine item quality. It suggests student-teachers improve their grammar mastery, and teacher education improves the quality of English grammar teaching and refine their curriculum of evaluation and language testing subjects and include the evaluation of student-teachers' performance in language assessment in teaching practicum program as a focus.

Article History
Received:
2 November 2023
Accepted:
26 October 2024
Available online:
9 November 2024

Keywords
item analysis; short-essay questions; formative test; pre-service English teachers; evaluation

Citation (APA Style): Hartati, N. (2024). Item analysis of pre-service EFL teachers' formative test in teaching practicum for program evaluation. *LingTera*, *11*(2), 145–158. https://doi.org/10.21831/lt.v11i2.67170

INTRODUCTION

Teaching practicum is essential for pre-service EFL teachers' professional development. It is an invaluable opportunity for student-teachers to implement the theoretical knowledge of teaching they have learned and put it into real practice in authentic classrooms. Wallace (1991) stated that teaching practicum provides student-teachers the opportunity to enhance their teaching professionalism which can only be attained through real practice in real classroom situations. In this program, the student-teachers are given an invaluable opportunity to apply the pedagogical content knowledge they have learned into real practice by teaching real students (Kosar, 2021). Thus, this program gives authentic hands-on experience of teaching English in real classrooms (Kim, 2020).

Teaching practicum is a mandatory subject in which student-teachers are sent to a certain school to take responsibility for teaching in classrooms. They are not only responsible for preparing the lesson plans and teaching the class but also for conducting an evaluation in which they have to construct a test to evaluate how much the teaching-learning activities and the students' learning outcomes have achieved the learning objectives. Conducting evaluation is an integral part of education that student-teachers must learn to be professional EFL teachers. Evaluation is a systematic process of collecting and analyzing information or data to make a judgment about a specific program

(EDC, 2013). By conducting an evaluation, they can make a judgement of the teaching-learning activities they are doing in the teaching practicum program, and get feedback on how far their teaching-learning activities and their students' learning outcomes have reached the learning objectives. By doing so they get information on what needs to be done in the next or future learning activities and what needs to be fixed and improved.

English Teachers' competencies in conducting evaluation and assessment in teaching-learning process are called Language Assessment Literacy (LAL) which is defined as a "basic understanding of sound assessment practice and the ability to apply that knowledge to measure language learning in different contexts" (Yan, et al., 2018, p. 158). It is the ability to determine what method of assessment to implement, how to assess, and when to assess to get information on students' learning performance (Jeong, 2013; Stiggins, 1999). Anam & Putri (2021), accommodating various literature on assessment literacy, summarized that LAL comprises five dimensions which are; designing test/assessment instruments, administering and scoring assessments, using alternative assessments, checking the validity and reliability of test instruments, and using assessment results to make decisions. Designing test instruments, thus, is one of the LAL that teachers must have to evaluate their teaching-learning process.

At schools, teachers commonly construct achievement tests to evaluate the teaching-learning process to measure students' learning outcomes that will enable them to measure students' knowledge, skills, ability, attitudes and performance (Brown, 2003), and also to inform teachers of the success of the teaching-learning activities that have been done in reaching the learning objectives. In a language classroom, an achievement test is usually designed to measure students' language and skill progress in relation to the syllabus being used to measure how well students have learnt what they have been studying and to determine what still needs to be done for further learning (Harmer, 2015).

There are two kinds of achievement tests that are usually administered by teachers at schools which are formative and summative tests. Ismail et al. (2022) explained that a summative test measures learning while a formative test allows for feedback which may improve learning. A summative test is a test that is usually administered at the end of a course (Liu et al., 2021; Rezai et al., 2022). It is done in order to make judgments about learning achievement, to measure how much students have learned from a learning course in a semester, and how much they have achieved the learning objectives (Brown, 2003; Houston & Thompson, 2017).

A formative test measures learners' abilities as part of a process and is part of the learning process itself. It looks to the future of what needs to be done to help students progress to the next level. Due to this reason, formative assessment is also called an assessment for learning or a progress test (Nicol & Macfarlane-Dick, 2006). Houston & Thompson (2017, p. 2) explained that "formative assessment was attached to improvement of learning in progress." It serves a feedback purpose to guide subsequent or future learning for students. Based on the pupils' performance in the formative test, teachers can make a decision on what needs to be done in future learning (Harmer, 2015). William (2011) explained that by conducting a formative test, teachers may get information about the teaching-learning process to make instructional decisions, as feedback for students to improve their performance that may motivate students to improve their learning activities. Formative tests are usually done frequently and interactively to assess students' understanding and development to detect their needs and to adjust teaching appropriately (Alahmadi, et al., 2019). It is an ongoing process that provides students constructive timely feedback to help them achieve the learning goals and enhance their achievements (Vogt et al., 2020). Thus, formative tests may be considered as the blending of assessment and teaching (Ozan & Kincal, 2018; Chan, 2021; Masita & Fitri, 2020). The essence of formative tests is to get information on students' advancement and to detect their major areas of weaknesses (Vadivel, et al., 2021).

Ismail, et al. (2022) found that formative tests are more effective than summative tests in improving students' academic motivation, test anxiety, and self-regulation skills. Further, they found that formative test is effective in helping students to "detect their own weaknesses and target areas that need more effort and work" (Ismail, et al., 2022, p. 1). Therefore, the practice of administering summative and formative tests in real classrooms in teaching practicum is an invaluable part of pre-service EFL teachers' professional development, especially in Language Assessment Literacy (LAL).

Considering the pivotal role of teaching practicum in EFL teacher training and education, a large amount of research has investigated this program in order to gain insights. Various studies have investigated pre-service English teachers’ performance in their teaching practicum program. Just to name a few, Lestari & Lestari (2022) investigated pre-service EFL teachers’ abilities in designing and implementing lesson plans. Astuti & Drajati (2022), Hendriwanto (2021), and Riyanti (2020) investigated pre-service English teachers’ self-reflections of their professional growth in teaching practicum. Hussein & Razeq (2022) and Nurjannah & Lestari (2021) investigated pre-service EFL teachers' experience conducting synchronous learning in teaching practicum during the coronavirus pandemic. Rachmawati, et al. (2017) examined pre-service EFL teachers’ self-concept from their reflection conducting their teaching practicum. Mahmoudi & Ozkan (2016) investigated various sources of stress faced by pre-service EFL teachers in teaching practicum and what kinds of strategies they used to cope with the stress. Eksi & Yakisik (2016) examined the reasons why pre-service EFL teachers experience or do not experience anxiety in reference to culture-specific reasons for conducting teaching practicum in Turkey. Rahayuningsih (2016) investigated challenges in developing teaching materials faced by pre-service EFL teachers in teaching practicum. Pasaka, et al. (2014) and Saricoban (2010) investigated challenges faced by pre-service EFL teachers in teaching practicum.

However, based on the writer’s literature review, research investigating pre-service EFL teachers’ performance in constructing an achievement test in their teaching practicum is still scarce. The writer could not find any research article investigating the quality of test items constructed by pre-service EFL teachers, let alone in their teaching practicum. In contrast, there is a large amount of research investigating the quality of test items, specifically on multiple-choice items, constructed by experienced EFL teachers in the Indonesian context such as by Hartati & Yogi (2019), Karim, et al. (2021), Darmawan et al., (2022), and also abroad such as by Kissi, et al. (2023), Lin (2018), and Toksoz & Ertunc (2017). This is in line with Anam & Putri’s (2021) literature finding that research investigating Language Assessment Literacy (LAL), including language construction ability, dominantly focused on in-service teachers’ performance both in Indonesian and abroad contexts. Due to this research gap, empirical evidence of pre-service EFL teachers’ competence in test construction is very limited (Anam & Putri, 2021), let alone in teaching practicum. Further, research investigating test items other than multiple-choice items, such as essay or short essay questions, constructed by both pre-service and in-service teachers is very limited.

To fill the research gap, therefore, this study intends to analyse the quality of formative test items in a short essay form, specifically in ‘completion’ or ‘fill-in-the-blank’ test items, constructed by pre-service EFL teachers in their teaching practicum program. Investigating pre-service EFL teachers’ performance in constructing a test in teaching practicum may give valuable insights into their competencies not only in test construction but also their teaching skills as well as their mastery of English in general which can give information and feedback for student-teachers, teaching practicum program and for teacher training and education institutions in general. This study intends to analyze the quality of a formative test constructed by two pre-service EFL teachers in their teaching practicum in terms of Facility Value (FV), Discriminating Power (DP), test reliability, and item validity as feedback for student-teachers and teacher training and education institutions. The researcher formulated the following two research questions as the research guidance:

1. How is the quality of the pre-service EFL teachers’ formative test items in terms of FV, DP, test reliability and item validity?
2. What can be learned from the quality of their formative test items for the student-teachers, teaching practicum program, and teacher training and education?

RESEARCH METHOD

This study used a qualitative and quantitative approach to analyze the data to answer the two research questions. To answer the first research question which is to analyze the quality of the test items, a quantitative item analysis was conducted to examine the Facility Value (FV) or the difficulty level, the Discriminating Power (DP), the test reliability and the item validity. Item analysis is a kind of document analysis that investigates the quality of test items by using the Item-response theory through the analysis of students' responses or answers to the test items. Item analysis is a process of analyzing the quality of a test instrument based on certain steps and procedures to sort out good items from bad ones that need to be eliminated or revised for future use (Musial et al., 2009). It is an analysis of the quality of each test item based on students' responses to each item. The purpose is to improve the quality of test items by identifying which items are good or bad and need to be revised or rejected for further use.

A qualitative and a quantitative approaches were utilized to analyze the quality of each test item to gain insights into the student-teachers' ability in test construction, the teaching-learning performance, and their mastery of English in general. The result of the quantitative and qualitative analyses were then triangulated to answer the second research question.

Research site and participants

This study involved two pre-service EFL teachers who were conducting their teaching practicum program in a private junior high school in Jakarta. As the prerequisites to joining the teaching practicum program, they had studied and passed the subjects of TEFL 1 and 2 (Teaching English as a Foreign Language) where they learned the development and history of Teaching Methodology, and TEFL 3 in which they had microteaching lesson of planning and delivering a lesson, and also Evaluation and Language Testing Development subject in which they learned how to construct a good test. They also studied and passed the three grammar subjects namely; Basic English Grammar, Intermediate English Grammar, and Advanced English Grammar. The two student-teachers were involved in this study because they were assigned by the supervisor-teachers of the school where they conducted their teaching practicum to construct the formative test items investigated in this study. There were five students conducting teaching practicum in this school and all of them were guided by two teacher-supervisors at the school who gave them advice, guidance and tasks in their teaching practicum. They were also supervised by a lecturer from the university where they studied.

Thirty-one students of an eighth-grade class taught by one of the pre-service EFL teachers who constructed the test items were involved. Their answers or responses to the test items of the formative test were then analyzed to investigate the quality of the test items.

Data Collection and Analysis

The formative test analyzed in this study consisted of 10 short essay questions in the form of ‘completion’ or ‘fill-in-the-blanks’ test items which used a dichotomy scoring system that had 10 scores for correct answers and 0 for wrong answers, but no score for half right or wrong. The test makers provided a clue of the answer for each question in a basic form and required students to answer in its correct form to complete each sentence based on its context. The ten points were only given to true answers that correspond to the exact form of answers in the anwer key. Any answer that does not fit the exact form was considered wrong and received 0 points. For example, if the answer key is 'bought' for question number 10, but the student's answer is 'baught', then it is considered as wrong with 0 points.

The questions were constructed based on the syllabus with five (50%) questions aimed to test students' mastery of the use of the degree of comparison, and 5 questions for Present Continuous tense.

1. Diego made chocolate … than Adi’s made. (Good)
2. Danny is the …boy in the class. (Clever)
3. Kenny is … than Kate. (clever)
4. The cost of living in Surabaya is … than in Jakarta. (Cheap)
5. I feel …than I did yesterday. (Happy)
6. They … lazy today. Do you see it? (Work)
7. Brandon and Rudi … football in the yard right now. (play)

8. We … in Tarakan City now. (Live)
9. They … bread in their kitchen right now. (Make)
10. Her mom … a vegetable in the market today. (Buy)

The students were given 15 minutes to do the test. After the test had been administered, the 31 students' answer sheets were then collected and scored. The students' responses and scores were then analyzed to determine the FV, DP, test reliability and item validity.

FINDINGS AND DISCUSSION

Findings

Quality of Pre-Service EFL Teachers' Formative Test Items: FV, DP, Reliability, and Validity

Facility Value (FV)

To determine the Facility Value (FV) or the level of difficulty of the essay question, Sudijono's (2012, p.134) formula was used. First, the average score of each item was calculated by summing up the total score achieved by all students for each question and then divided by the total number of students. To calculate the level of difficulty, the average score of each item was then divided by the maximum score of each item which was then determined by its qualification by consulting (Arikunto 2018) criterion of Facility Value.

$$\text{Average score for each question} = \frac{\text{The total score of all students for each question}}{\text{The total number of students}} \quad (1)$$

$$\text{FV} = \frac{\text{The average score of each item}}{\text{Maximum score of each item}} \quad (2)$$

To determine the qualification of the difficulty level, Arikunto's (2018, p.225) criterion was consulted.

0.71 – 1 .00 = Easy
0.31 – 0.70 = Medium
0.00 < 0.30 = Difficult

For example, for question number 1, the total score of all students was 260. To get the average score is by dividing 260 by 31 of the total number of students which is 8.39. The Facility Value is 0.839 gained by dividing 8.39 by 10 as the maximum score for each question. Based on Arikunto's (2018, p.225) classification of the difficulty level, question number 1 with the score of FV level 0.839 belongs to an easy question. The result of the analysis of the Facility Value or the difficulty level of each question is shown in Table 1.

Table 1. The Result of Facility Value Analysis

Difficulty Category	Number of Items	Number of Questions	Percentage
Easy	1, 4, 5, 9	4	40%
Medium	2, 3, 6, 7, 8, 10	6	60%

As shown in the table, the difficulty level of the short-essay questions only consisted of easy and medium levels with a ratio of 2:3 but there is no difficult question. According to Sumarsono (2000), a good test should contain 25% for both easy and difficult questions, and 50 % for medium questions. However, the pre-service EFL teachers did not construct difficult questions which indicates the pre-service EFL teachers' inability to construct difficult questions or higher-order thinking skill questions.

Discriminating Power (DP)

Discriminating Power (DP) is the ability of each item to distinguish the students who have mastered the tested material from the students who have not. The index of the DP ranges from 0.00 to 1.00. The higher the index, the better the ability of the test item to differentiate the students who have mastered the learning material from those who have not.

To analyze the Discriminating Power (DP) of each test item, the students were grouped into the upper-group level (UG) and the lower-group level (LG) based on their total correct answer to the test. Students who got scores ranging from 0 - 50 were considered in the LG, while the students whose scores ranged from 60-100 were put in the UG. There were 9 students in the LG and 21 students in the UG. The formula used to analyze the Discriminating Power (DP) was adapted from Arikunto (2003, p. 238).

$$DP = \frac{Gu}{U} - \frac{Gl}{L} \qquad (3)$$

Gu = The number of the upper-level students (UG) who answered the item correctly
U = The total number of students in the upper-level group (UG)
Gl = The number of the lower-level (LG) students who answered the item correctly
L = The total number of students in the lower-level group (LG)

Arikunto's (2003, p. 232) classification of the score of the DP was used.

0.70 – 1.00	= excellent
0.40 – 0.69	= good
0.20 – 0.39	= satisfactory
0.00 – 0.19	= poor

For example, the Gu or the number of students in the UG who answered question number 1 correctly was 20 of the total 21 UG students, and the Gl or the number of students in the LG who answered question number 1 correctly was 5 out of 9 students. The Discriminating Power (DP) of question number 1 is:

$$DP = \frac{20}{21} - \frac{5}{9} = 0.4$$

Table 2. The Result of Discriminating Power Analysis

DP Category	Number of Items	Number of Questions	Percentage
Excellent	7, 8, 9	3	30%
Good	1, 2, 5, 6, 10	5	50%
Satisfactory	3, 4	2	20%

All of the 10 items have excellent, good, and satisfactory levels of Discriminating Power (DP) which means that all items can differentiate students' mastery of the learning materials with 30% of excellent quality, 50% in medium quality, and 20% in satisfactory quality.

Test Reliability

Reliability of the test refers to the level of consistency of an instrument that is whether it measures consistently when being tested and re-tested to the same subjects or test takers. In other words, a test is reliable if it always gives the same result when tested and re-tested in the same group of students at a different time (Arifin 2012, p. 258).

In this study, to test the reliability of the test, the split-half method of reliability test was employed. The items were divided into two halves which consisted of the odd number items and the even number items. The scores of the first half are considered as variable x and the scores of the other half are considered as variable y. The formula of Pearson Product Moment Correlation was employed to get the Coefficient Correlation of the half items (rgg).

$$r_{gg} \text{ or the } r_{xy} = \frac{N\sum xy - (\sum x)\,(\sum y)}{\{N\sum x2 - (\sum x)\}\,\{N\sum y2 - (\sum y)2\}} \qquad (4)$$

The score of the r_{gg} was 0.819216. To calculate the reliability of the whole test, the Brown formula was then used:

$$r_{tt} = \frac{2 \text{ x } r_{gg}}{1 \text{ x } r_{gg}} \qquad (5)$$

Notes:
r_{tt} = Coefficient reliability of a test
r_{gg} = even and odd Correlation Coefficient (half the test with the other half)

Using the formula of Spearman-Brown above, the rtt was then gained with the result of 0.900625. According to Sudijono, (2011, p. 209), if the rtt > 0.7, it means the test has a high reliability but if rtt < 0.7, the test is unreliable. This finding indicates that the test has a high reliability.

Item Validity

Each item of a test is an integral part of the whole test, and the validity of each item contributes to the validity of a test as a whole. In other words, the validity of a test as a whole depends on the validity of each item (Sudijono 2012). Therefore, to check the validity of a test can be analyzed by analysing the validity of each test item. Item validity is the accurate measurement of each item in measuring what is intended to measure (Sudijono 2012).

Because this formative test used a dichotomy scoring, which was 0 for a wrong answer and 10 for a correct answer, to analyze the validity of each item, the Point Biserial Correlation formula was used (Sudijono 2012, p. 185).

$$\gamma_{pbi} = \frac{M_p - M_t}{S_t}\sqrt{\frac{p}{q}} \qquad (6)$$

γ_{pbi} = Biserial Correlation Coefficient
M_p = Mean score of the students who responded correctly to the analysed item
M_t = Mean of the total score of all students
S_t = Standard Deviation of the total score
p = The proportion of students who answered the item correctly
q = The proportion of students who answered the item incorrectly (q = 1-p)

The gained Point Biserial Correlation index (γpbi) was then consulted with the r table at the level of significance of 5% based on the number of students involved in this study (Sudijono 2012, p. 185).

Table 3 indicates that there were only 4 (40%) valid items, while the rest 6 (60%) are invalid. According to Sudijono (2012), the invalidity or the low score of validity of test items is an indicator that there is something wrong with the test and that the test makers should be cautious. It is an indicator that the test items failed to measure what are supposed to measure. Sudijono (2012) further

Table 3. The Result of Each Item Validity Analysis

Category	Number of items	Number of questions	Percentage
Valid	1, 2, 4, 6	4	40%
Invalid	3, 5, 7, 8, 9, 10	6	60%

explained that the low validity or the invalidity of each test item is caused by the number of test takers who cannot answer the item correctly. Sudijono's (2012) formula to test item validity above also indicates that the higher the number of students who can answer each item correctly, the higher the score of the validity score. In other words, the more students who master the learning material being tested in the item, the higher the score of the validity of the test item. This discussion suggests that the invalidity of the six items indicates that the students have not mastered the materials being tested in those six items.

Insights from Formative Test Item Quality for Student-Teachers, Practicum Programs, and Teacher Training

To gain a deeper understanding of the quality of the test items and gain feedback on the student-teachers' ability in test construction, the teaching-learning performance, and their English mastery, both the test items and the students' answers were then analyzed qualitatively and quantitatively. This second step of analysis was done to get more insight on the quality of each test item to answer the second research question.

The analysis of the students' responses to question number 3 on the degree of comparison revealed that 15 (48%) students have not mastered the form of the comparative degree of the word 'clever'. Those students used 'most clever', 'cleavernest', 'clevernes', 'more cleverest', 'clever', 'more celeverer', 'more cleves', 'cleveries', 'most cleverer' rather than 'more clever' or 'cleverer'. As stated earlier, these answers are incorrect because they do not match the exact answer key, and thus, it received 0 points. A similar finding was also found for question number 5. Six (19%) students failed to give the correct answer as they used 'happyer', 'more happy', 'more happier', 'happying', rather than 'happier'. This finding indicates that there is a need for remedial teaching for the comparative degree of the adjective 'clever' and for any adjective with -y endings.

For question number 7 for Present Continuous Tense, 14 (45%) students failed to answer this item correctly. Ten (32%) students did not use 'to be', and 4 (12%) used to be 'is' for the subjects 'Brandon and Rudi'. This finding indicates that they have not mastered the structure of the Present Continuous Tense, specifically on the use of 'To be'. Further, the use of to be 'is', rather than 'are' also indicates that the students have not mastered the use of plural 'to be' for plural collected subject. A similar finding for question number 9 revealed that 9 (29%) students failed to give the correct answer because they missed the use of 'to be' but used the correct form of the verb 'making', and 1 student used the word 'maker'.

A similar finding was also found for question number 8 which was meant to test the use of Present Continuous Tense. Ten (32%) students did get the score for this item. Eight students did not use 'To be' but used the correct verb form of 'living', and two students did not use 'To be' and used the incorrect verbs 'leaving' and 'liveaving' which indicates that these 2 students have not known the -ing form of the word 'live'. This analysis informs teachers that they need to re-teach the formula of Present Continuous Tense and the -ing forms of the verbs.

However, a grammatical analysis of item 8 "We … in Tarakan City now. (Live)" revealed its grammatical error. This item was intended to test the students' mastery of Present Continuous Tense. However, the use of the verb "Live" for temporal action is inappropriate because it is usually used to refer to general truth (Simple Present Tense), but not for temporal action. The word "stay" is more appropriate in this context. Two students used Simple Present Tense to answer this item although they used the incorrect verb "lives" for the subject "We". Thus, this item failed to measure, at least, the two students' mastery of the use of Present Continuous Tense.

The same finding was also found for question number 10. Eight (26%) students did not get the score. The reason for this was because 4 (13%) students did not use ‘to be’ and used the incorrect -ing form of the word ‘buy’ of which they used ‘buyying’ and ‘buyer’. One of them also used to be ‘are’ for the subject ‘Her mom’. However, the rest 4 (13%) of the students did not get the score caused by the grammatical error of this item. Question number 10. “Her mom … a vegetable in the market today. (Buy)” is ambiguous in the use of Tenses caused by the ambiguous time signal ‘today’. Although the student-teachers intended to test the students’ knowledge of the use of Present Continuous Tense, this sentence may mean that the event of “her mom” bought a vegetable happened in the earlier time of ‘today’ (Past Tense). Therefore, this sentence may be answered in Simple Past Tense. Thus, the time signal ‘today’ should be replaced by the word ‘now’ to avoid confusion. Because of this confusion, four students answered it in Simple Past tense. Thus, this item failed to measure what was intended to measure, at least for those 4 students, whether they have mastered the use of Present Continuous Tense or not.

Further, there is another question with grammatical error which is question number 1 "Diego made chocolate …than Adi's made. (Good)". The question is confusing in terms of meaning. It was ambiguous what was being compared whether the quality of chocolate or the way Diego and Adi made chocolate. Because of this, four students used 'to be + the comparative degree of good' ("is better"). Meanwhile, the test makers intended to compare the verb or the way Diego and Adi made chocolate. Therefore, the question should be revised into "Diego made chocolate … than Adi did. (Good)". However, the student-teachers seemed to realize the grammatical error so they considered those four students' answers correct. Because of this decision, it did not affect the score validity of this item.

The grammatical errors of those three items, to some extent, influenced the score validity of each of those three items, the DP, and the FV which indicates that the test maker’s English grammar mastery determines the quality of a test item and a test as a whole.

Discussion

The formative test gives valuable feedback for the pre-service EFL teachers to measure how much their students and the teaching-learning process have reached the learning objectives. The item analysis of the 10 short essay questions revealed which part of the learning materials have not been mastered by the students and what should be done in future teaching-learning activities.

The item analysis also revealed the pre-service EFL teachers’ ability in constructing test items and their English grammar mastery. The analysis of Discriminating Power (DP) revealed that the items have good level which means that the items can discriminate students who have mastered the learning materials from those who have not. The formative test reliability is high which indicates its good consistency in measurement. However, the analysis of the item validity revealed that 60% of the items are not valid caused by a small number of students who could answer the items correctly. This indicates that the pre-service EFL teachers should give remedial teachings for the aspects that have not been mastered by the students.

Further analysis revealed that three items contain grammatical errors which caused ambiguity and confusion for their students in answering the questions which, in turn, influenced the Facility Value, the Discriminating Power, and the item validity of those three items. The grammatical errors were due to the pre-service EFL teachers' lack of mastery of English tenses and the grammatical aspects being tested. This suggests the need for the pre-service EFL teachers to improve the mastery of English Tenses and English use in general which also informs teacher training and education to improve the quality of their education, especially in the teaching of English grammar.

The analysis of the Facility Value (FV) or the difficulty level revealed that there was no difficult question found in the test which indicates that the student-teachers have not been able to make difficult test items. This finding, further, indicates that the pre-service EFL teachers have not been able to implement Bloom’s taxonomy and Higher Order Thinking Skills (HOTS). Of the six levels of higher order thinking skills; knowledge and recall, comprehension and understanding, application and context, analysis, synthesis, and evaluation (CETL, 2024), the test items only focused on testing students' first level of higher order thinking skills which was on knowledge and recall of adjective forms for degree of comparisons and the verb pattern of Present Continuous Tense.

Although the writer did not find any research article investigating the same area to compare their performance, a numerous item analysis researching experienced or in-service teachers' constructed multiple-choice test items revealed that they also produced a limited number of difficult questions compared to easy and moderate questions. For example, Hartati & Yogi's (2014) research on multiple-choice item analysis of a summative test constructed by experienced teachers in a senior high school in Indonesia revealed that the proportion of difficult questions was far less in number (12% or 6 out of 50 questions). Karim, et al.'s (2021) item analysis on English multiple-choice test items constructed by an in-service teacher in Indonesia also revealed less in the percentage of difficult questions that out of 50 questions, there was only 1 (0.5%) difficult question found. Darmawan, et al.'s (2022) research investigating the quality of English multiple-choice test items constructed by an in-service teacher in Indonesia also revealed a similar finding that the proportion of difficult questions is far less in number (4 or 10% out of 40 questions). The findings indicate that even experienced or in-service teachers are still less competent in constructing difficult questions which means that even in-service or experienced teachers still have difficulty in applying Bloom's Taxonomy or Higher Order Thinking Skills (HOTS) in test construction. This finding is in line with various research findings that even in-service second or foreign language teachers still lack language assessment literacy including in test construction (Darmawan, et al., 2022; Anam & Putri, 2021; Karim, et al., 2021; Umam & Indah, 2020; Lam, 2019; Hartati & Yogi, 2019; Tsagari & Vogt, 2017; Nemati, et al., 2017; Qian, 2014; Popham, 2001; Brookhart, 2001). Anam & Putri's (2021) research investigating how pre-service and in-service English language teachers in Indonesia self-rated their Language Assessment Literacy (LAL) revealed that they admitted that their test construction ability is still at a moderate level and they admitted to not having the knowledge to conduct item analysis to check the validity and reliability of their test items. The reason for this, according to Schafer & Lissitz (1987) and Schafer (1993), may be due to inadequate preparation or learning of assessment in their previous teacher education program. This is also supported by Wise, et al.'s (1991) research finding in which in-service teachers reported that they felt that they did not receive adequate training in assessment. From the discussion of the previous studies above, it can be concluded that the reason for the pre-service EFL teachers' difficulty in constructing test items, specifically difficult test items, may be due to inadequate learning or practice in test construction in evaluation and language testing development subject in the teacher education program. Anam & Putri (2021) also shared the same opinion in response to their finding that both pre-service and in-service teachers' low self-rating in LAL, especially in test construction and item analysis, may be due to the courses in their teacher education have not given ample opportunities for them to develop their LAL.

The teaching practicum program is a good opportunity for pre-service EFL teachers to learn and have an actual practice of conducting evaluation and assessment in real classrooms with supervision, guidance, and feedback from their supervisor-teachers at the school of the teaching practicum program. Anam & Putri's (2021) research found that despite in-service teachers' Language Assessment Literacy (LAL) is still at a moderate level, their self-rating in LAL is higher than that of the pre-service EFL teachers in designing test instruments, administering and scoring assessments, using alternative assessment, and using the test results to make decisions. Therefore, the curriculum for teaching practicum should also make evaluation and language test development a focus and make it one of the components of evaluation and assessment of their performance in teaching practicum. By doing so, the supervisor-teachers will pay more attention and give their guidance and feedback to the student-teachers in administering evaluation and assessments. This will improve student-teachers' LAL in administering evaluation and assessments for teaching and learning. From the writer's observation of teaching practicum, the supervision still focuses more on other aspects of teaching such as lesson planning, classroom management, and teaching material development, however, test construction has not been well-paid attention.

CONCLUSION

The analysis of the quality of the test items constructed by pre-service EFL teachers in teaching practicum has given invaluable insights and feedback not only for the student-teachers but also for teaching practicum program and English teacher education in general. For pre-service EFL teachers, it is important to always improve their English grammar mastery as it is one of the prerequisites for teaching and their grammar mastery also influences the quality of their test items. Further, they also have to improve their Language Assessment Literacy (LAL) for their professional development.

The findings and discussion of the study also revealed important feedback for the improvement of the curriculum of teaching practicum program and teacher education. For the improvement of teaching practicum program, it is crucial to include student-teachers' performance in test administration in teaching practicum as a focus to enable the student-teachers to get guidance and feedback from the in-service supervisor-teachers at the school. This will enable student-teachers to learn and improve their professional development in Language Assessment Literacy (LAL) and other aspects of teaching as research has shown that in-service teachers' Language Assessment Literacy (LAL) is higher than that of pre-service EFL teachers.

The findings and discussion above also give invaluable feedback for the improvement of English teacher training and education in general. It informs the teacher training and education how much their students are prepared to enter their teaching profession, their pedagogic skills, their Language Assessment Literacy (LAL), and even their mastery of English grammar and use in particular. It gives feedback for teacher education to improve the quality of English grammar teaching. Further, the pre-service EFL teachers' lack of competencies in test construction, such as their difficulty in constructing difficult questions in particular, which literature has shown is also experienced by in-service English teachers, suggests English teacher training and education institutions pay more attention to the refinement of their curriculum for the evaluation and language testing development subject where the student-teachers should be facilitated to have adequate training and practice in test construction. Further, it is also important to equip the pre-service EFL teachers with the skill and knowledge in conducting item analysis to check and analyze the quality of their test item for their professional development.

Furthermore, due to the scarcity of research investigating pre-service English teachers' performance in administering evaluation and assessment, specifically in test construction in teaching practicum program, some suggestions are offered for future research. First, future research should also conduct interviews with pre-service EFL teachers on administering tests in their teaching practicum to get more insight into the process and their follow-up actions regarding the result of the administered formative test. Second, future research may investigate pre-service EFL teachers' ability in constructing multiple-choice items in formative and summative tests, and other test item formats to fill the research gap.

ACKNOWLEDGEMENT

This work was supported by the Department of Research and Development or *Lembaga Penelitian dan Pengembangan* Universitas Muhammadiyah Prof. DR. Hamka under an Internal grant with an SPK number. 742/F.03.07/2022.

REFERENCES

Alahmadi, N., Alrahaili, M., & Alshraideh, D. (2019). The Impact of the Formative Assessment in Speaking Test on Saudi Students' Performance. *Arab World English Journal, 10(1)*, 259–270.

Anam, S., & Putri, N. V. (2021). How Literate am I about Assessment: Evidence from Indonesian EFL Pre-service and In-service Teachers. *ENGLISH REVIEW: Journal of English Education, Volume 9, Issue 2, June*, 377-388.

Arifin, Z. (2012). *Evaluasi Pembelajaran.* PT. Remaja Rosdakarya.

Arikunto, S. (2018). *Dasar-dasar Evaluasi Pendidikan, edisi 3.* Bumi Aksara.

Astuti, Y. D., & Drajati, N. A. (2022). Teaching Practicum Experiences: Pre-service English Teachers' Self-Reflections of Their Professional Growth. *Journal of Innovation in Educational and Cultural Research, Vol 3 (3)*, 382-389.

Brookhart, S. M. (2001). *The standards and Classroom Assessment research. In Paper presented at the annual meeting of the American Association of Colleges for Teacher Education (p. 15).* ERIC Document Reproduction Service No. ED451189.

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

CETL. (2024, November 6). *Critical Thinking and other Higher-Order Thinking Skills*. Retrieved from Center for Excellence in Teaching and Learning: https://cetl.uconn.edu/resources/design-your-course/teaching-and-learning-techniques/critical-thinking-and-other-higher-order-thinking-skills/

Chan, K. T. (2021). Embedding Formative Assessment in Blended Learning Environment: The Case of Secondary Chinese Language Teaching in Singapore. *Education Sciences, 11(360)*, 1-12.

Darmawan, M., Sudarsono, Riyanti, D., Yullana, Y. G., & Sumarni. (2022). A Test Items Analysis of English Teacher Made Test. *Journal of English Education and Teaching (JEET), Volume 6, number 4*, 498-513.

EDC. (2013, September 22). *Understanding Evaluation: promote Prevent-3 Bold Steps*. Retrieved from Education Development Centre, Inc: http://positiveschooldiscipline.promoteprevent.org/

Eksi, G. Y., & Yakisik, B. (2016). To Be Anxious or Not: Student Teachers in the Practicum. *Universal Journal of Educational Research (46)*, 1332-1339.

Harmer, J. (2015). *The Practice of English Language Teaching 5th edition.* Pearson.

Hartati, N., & Yogi, H. P. (2019). Item Analysis for a Better Quality Test. *English Language in Focus (ELIF)*, 59-69.

Hendriwanto. (2021). A Reflective Teaching Practicum as a Platform for Stimulating Teachers' Professional Development. *Journal of Education for Teaching*, 1-3.

Houston, D., & Thompson, J. N. (2017). Blending Formative and Summative Assessment in a Capstone: 'It's not your tools, it's how you use them'. *Journal of University Teaching & Learning Practice, Vol. 14, Issue 3,* , 1-13, http://ro.uow.edu.au/jutlp/vol14/iss3/2.

Hussein, A., & Razeq, A. (2022). Teachers Candidates' Experience in a Practicum for English as a Foreign Language. *International Journal on Studies in English Language and Literature (IJSELL), Vol. 11, Issue 2*, 2347-3134.

Ismail, S., Rahul, D. R., Patra, I., & Rezvani, E. (2022). Formative vs. Summative Assessment: Impacts on Academic Motivation, Attitude toward Learning, Test Anxiety, and Self-regulation Skill. *Language Testing in Asia*, 1-23, https://doi.org/10.1186/s40468-022-00191-4.

Jeong, H. (2013). Defining Assessment Literacy: Is it different for language testers and non-language testers? *Language Testing, 30(3)*, 345-362. https://doi.org/10.1177/0265532213480334.

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*, 256-269.

Kim, J. (2020). Learning and Teaching Online During Covid-19: Experiences of Student Teachers in an Early Childhood Education Practicum. *International Journal of Early Childhood 52(2)*, 145-158.

Kissi, P., Baidoo-Au, D., & Anane, E. (2023). Teachers' Test Construction Competencies in Examination-oriented Educational System: Exploring Teachers' Multiple-choice Test Construction Competence. *Frontiers in Education, Volume 8*, 1-14. https://doi.org/10.3389/feduc.2023.1154592.

Kosar, G. (2021). Distance Teaching Practicum: Its Impact on Pre-service EFL Teachers' Preparedness for Teaching. *IAFOR Journal of Education, vol 9 (2)*, 111-126.

Lam, R. (2019). Teacher Assessment Literacy: Surveying knowledge, conceptions and practices of classroom-based writing assessment in Hong Kong. *System, 81(1)*, 78-89. https://doi.org/10.1016/j.system.2019.01.006.

Lestari, A. E., & Lestari, S. (2022). Pre-service English Teachers' Practices of Designing and Implementing Lesson Plans for Teaching Practicum. *LingTera, 9(1)*, 25-36.

Lin, S. (2018). Item Analysis of English Grammar Achievement Test. *Mandalay University of Foreign Languages Research Journal, Vol. 9, No. 1*, 13-20.

Liu, F., Vadivel, B., Rezvani, E., & Namaziandost, E. (2021). Using Games to Promote EFL Learners' Willingness to Communicate (WTC): Potential Effects and Teachers' Attitude in Focus. *Frontiers in Psychology, 12 (762447)*, 1-10.

Mahmoudi, F., & Ozkan, Y. (2016). Practicum Stress and Coping Strategies of Pre-service English Language Teachers. *Procedia - Social and Behavioral Sciences 232: International Conference on Teaching and Learning English as an Additional Language*, 494-501.

Masita, M., & Fitri, N. (2020). The Use of Plickers for Formative Assessment of Vocabulary Mastery. *Ethical Lingua Journal of Language Teaching and Literature, 7(2)*, 311–320.

Musial, D., Nieminen, G., Thomas, J., & Burke, K. (2009). *Foundations of Meaningful Educational Assessment.* McGraw-Hill Higher Education.

Nemati, M., Alavi, S. M., Mohebbi, H., & Masjedlou, A. P. (2017). Teachers' Writing Proficiency and Assessment Ability: The Missing Link in Teachers' Written Corrective Feedback Practice in an Iranian EFL Context. *Language Testing in Asia, 7(1). https://doi.org/10.1186/s40468-017-0053-0*.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative Assessment and Self-regulated Learning: A Model and Seven Principles of Good Feedback Practice. *Studies in Higher education, Vo. 31. No. 2, April*, 199-218.

Nurjannah, & Lestari, S. (2021). The Teaching Practicum Experience of Pre-service English Language Teachers through Synchronous Learning. *EDULINK (Education and Linguistics Knowledge Journal), vol. 3. (2)*, 92-115.

Ozan, C., & Kıncal, R. Y. (2018). The Effects of Formative Assessment on Academic Achievement, Attitudes toward the Lesson, and Self-regulation Skills. *Educational Sciences: Theory and Practice, 18*, 85-118.

Pasaka, R. W., Emilia, E., & Amalia, L. L. (2014). The Exploration of Pre-service EFL Teachers' Challenges in Field Practicum. *International Journal of Education, Vol. 7, No. 2, May*, 144-153.

Popham, W. J. (2001). *The Truth about Testing: An educator's call to action.* ASCD (Association for Supervision and Curriculum Development).

Qian, D. D. (2014). School-based English Language Assessment as a High-stakes Examination Component in Hong Kong: Insights of Frontline Assessors. *Assessment in Education: Principles, Policy & Practice, 21(3)*, 251-270. https://doi.org/10.1080/0969594X.2014.915207.

Rachmawati, D., Emilia, E., & Lukmana, I. (2017). Self-concept of EFL Pre-service Teachers: A Reflection from a Teacher Practicum in Indonesia Context. *The Journal of English Language Studies, Vol. 02, No. 01,*, 1-18.

Rahayuningsih, D. (2016). Student Teachers' Challenges in Developing Teaching Materials During Teaching Practicum in Vocational School. *Journal of English and Education 4(2)*, 24-34.

Rezai, A., Namaziandost, E., Miri, M., & Kumar, T. (2022). Demographic Biases and Assessment Fairness in Classroom: Insights from Iranian University Teachers. *Language Testing in Asia, 12(1)*, 1–20. https://doi.org/10.1186/s40468-022-00157-6.

Riyanti, D. (2020). Students' Reflections in Teaching Practicum: A case study of EFL Pre-service Teachers. *Journal on English as a Foreign Language, Vol. 10 (20)*, 268-289.

Saricoban, A. (2010). Problems Encountered by Student-teachers during Their Practicum Studies. *Procedia Social and Behavioral Sciences 2*, 707-711.

Schafer, W. D. (1993). Assessment in Teacher Education. *Theory into Practice, 32(2)*, 118-126.

Schafer, W. D., & Lissitz, R. W. (1987). Measurement Training for School Personnel: Recommendations and reality. *Journal of Teacher Education, 38(3),* 57-63.

Stiggins, R. J. (1999). Assessment, Student Confidence, and School Success. *Phi Delta Kappan, 81(3),* 191-198.

Sudijono, A. (2012). *Pengantar Evaluasi Pendidikan.* Rajawali Press.

Sumarsono, S. (2020). *Metode Penelitian.* UHAMKA.

Toksoz, S., & Ertunc, A. (2017). Item Analysis of a Multiple-Choice Exam. *Advances in Language and Literacy Studies, Volume 8, Issue 6*, 141-146. http://dx.doi.org/10.7575/aiac.alls.v.8n.6p.140.

Tsagari, D., & Vogt, K. (2017). Assessment Literacy of Foreign Language Teachers around Europe: Research, challenges and future prospects. *Language Testing and Assessment, 6(1),* 41-61.

Umam, A., & Indah, Y. A. (2020). Exploring In-Service TEYL Teachers' Assessment Literacy: Implication for Continuing Professional Development. *Journal of English Educators Society, 5(1),* 47-51. https://doi.org/10.21070/jees.v5il.364.

Vadivel, B., Namaziandost, E., & Saeedian, A. (2021). Progress in English Language Teaching through Continuous Professional Development—Teachers' Self-Awareness, Perception, and Feedback. *Frontiers in Education, 6, 757285,* 1-10, https://doi.org/10.3389/feduc.2021.757285.

Vogt, K., Tsagari, D., & Csépes, I. (2020). Linking Learners' Perspectives on Language Assessment Practices to Teachers' Assessment Literacy Enhancement (TALE): Insights from Four European Countries. *Language Assessment Quarterly,* 1-24, https://doi.org/10.1080/15434303.2020.1776714.

Wallace, M. J. (1991). *Training Foreign Language Teachers: A Reflective Approach.* Cambridge University Press.

William, D. (2011). *Embedded Formative Assessment.* Solution Tree Press.

Wise, S. L., Lukin, L., & Roos, L. (1991). Teacher Beliefs about Training in Testing and Measurement. *Journal of Teacher Education, 42(1),* 37-42.

Yan, X., Zhang, C., & Fan, J. J. (2018). "Assessment knowledge is important, but ...": How contextual and experiential factors mediate assessment practice and training needs of language teachers. *System, 74,* 158-168. https://doi.org/10.1016/j.system.2018.03.003.