TEACHERS' ASSESSMENT LITERACY IN APPLYING PRINCIPLES OF LANGUAGE ASSESSMENT

Diah Wulansari Hudaya

English Education Department, Universitas Pendidikan Indonesia
diahwulanhudaya@student.upi.edu

Abstract

Assessment is an essential activity in teaching and learning process. However, most teachers do not feel adequately prepared in assessing their students. Therefore, this study focuses on the investigation of the teachers' assessment literacy in terms of teachers' preparation level in assessing students' performance, teachers' practices in applying principles of language assessment in their classroom, and the usefulness of the questionnaire developed for them. Forty three in-service teachers were administered 31 items questionnaire that covers principles of language assessment with 1-4 Likert scale, 3 open-ended questions and 7 items of background questionnaire. The quantitative data were analyzed by means of descriptive statistics in terms of percentage, mean, and standard deviation and the qualitative result were summarized, categorized, and analyzed for frequency of mention. The results showed that 79 % teachers felt prepared in assessing students' performance, it was supported teachers' practices in applying the principles (88.7 % agree), and 86 % teachers thought that the questionnaire that covers the principles of language assessment was useful for them to evaluate and/ or design a test. Finally, it is recommended that by applying the principles of language assessment in making a good test, teachers will be able to improve their instruction and understand how to use the test to help their students in improving their success in the target language.

Keywords: assessment literacy, principles of language assessment, teachers' practices

Introduction

Assessing student performance is one of the most critical aspects of a classroom teacher (Mertler, 2009). Regrettably, although most of teachers completed their teacher-education program, there was no requirement that they learn anything from their education program (Popham, 2009). Furthermore, Stiggins (2002) observed that teachers rarely have the opportunity to learn how to use assessment as a teaching and learning tool. On the other hand, without a higher level of teacher assessment literacy, teacher will be unable to help students attain higher levels of academic achievement (Coombe, Troudi, & Al-Hamly, 2012).

One of the assessment devices is test. Tests are a subset of assessment, although they are not the only form of assessment, but tests can be useful devices to assess student (Brown, 2004). Moreover, most classroom assessment involves tests that teachers have constructed themselves (Frey & Schmitt, 2007). However, it is unfortunate that learners all too often view tests as dark clouds hanging over their heads, upsetting them with anxiety as they get through questions they cannot answer and the worst is disappointment if they don't have good grade (Brown, 2001). Regarding the issue, then Stiggins (2002) stated that it is better to first ask, how can we use assessment to

help all our students want to learn and how can we help them feel able to learn? From the question, it is important that teacher should be able to design a good test and evaluate the existing test, not only to help students, but also to make a test become positive experience for students. Furthermore, Brindley (2001) recognized that teachers work within time and resource constraints and urged testing educators to recognize that they must develop flexible approaches to their assessment practices. Therefore, it is crucial to investigate teachers' assessment literacy in terms of their preparation level in assessing students' performance and their practices in applying principles of language assessment in designing a good test and evaluating the existing test.

Assessment literacy is defined as the understanding of (1) the means for assessing what students know and can do, (2) the interpretation of the results from these assessment, and (3) application of assessment results to improve student learning and program effectiveness (Webb, 2002). According to Boyle (2006), assessment literacy is simply an understanding of the principles and practices of testing and assessment. However, Mertler (2003) stated that many teachers believe that the assessment training that they received as undergraduates did not prepare them to be comfortable with the decisions they are routinely charged to make. Furthermore, Stiggins (2014) asserted that teachers spend as much as a quarter to a third of their professional time involved in assessment-related work without the essential training to do it well.

Related researches about assessment literacy have been conducted for years. Mertler (2003) conducted a research on pre-service and in-service teachers using the survey of Classroom Assessment Literacy Inventory (CALI) adapted from the similar instrument called *Teacher Assessment Literacy Questionnaire* which was developed from *The Standards for Teacher Competence in the Educational Assessment of Students*. It was found that the in-service teachers' level of assessment literacy was higher than pre-service teachers', but the two groups performed poorly in developing valid grading procedures. Moreover, Mertler's (2009) study examined the effectiveness of a two-week classroom assessment workshop for in-service teachers which based on The Standards for Teacher Competence in the Educational Assessment of Students. The result showed that the post-test was higher that the pre-test.

Mertler and Campbell (2005) conducted a research to pre-service teachers using Assessment Literacy Inventory (ALI). They found that their score was far lower than might otherwise be expected given their recent completion of coursework in classroom assessment. The reason they elicited was perhaps because ALI was designed to measure real-world application concept and competencies in The Standard, limited familiarity and experience with day to day realities in the classroom may have precluded pre-service teachers from making necessary connection. Volante and Fazio (2007) found that teacher candidates' self-efficacy of assessment literacy across each of the four years of their undergraduate program remained relatively low and they use the assessment primarily for traditional summative purposes. DeLuca & Bellara (2013) found that beginning teachers continue to maintain low competency levels in assessment. Their findings are similar to what is found by Yamtim & Wongwanich (2014). They asserted that most of the participants in their study, primary teachers, had scores for classroom assessment literacy at the poor level. Thus, the finding indicates that teachers do not possess one essential capacity necessary to fulfill their duties, which is to conduct learning assessment. What's more surprising, Perry (2013)

suggested that the level of teachers' assessment literacy on CALI have not changed significantly in over twenty years.

Crusan, Plakans, Gebril (2016) stated interesting finding in their study. In the contrary with the previous study, such as study by Mertler (2003 & 2009), Mertler & Campbell (2005), Perry (2013), and Yamtim & Wongwanich (2014), their survey showed more positive picture with regard to assessment, especially in writing assessment. Another interesting finding is that more experienced teachers, reported less assessment knowledge and less experienced teachers reported higher assessment knowledge than teachers with 11-20 years of teaching experience. Although the reason of this was unclear, they suggested that it may be caused by the changes over time in pre-service teacher education or the presence or absence of in-service training. The last, Newfields (2006) suggested that assessment literacy is an important aspect of overall teacher development and teachers who want to develop professionally should learn more about assessment.

Beside the result of the previous researches above, there are some obstacles called barriers to assessment literacy which prevent the development of assessment literacy. (1) Fear represents a prominent barrier to assessment literacy because it closes many educators off from even reviewing their own assessment competence (Stiggins, 1995). (2) There are insufficient resources allocated to assessment, such as journals and books (Coombe, Troudy, & Al-Hamly, 2012). (3) Teachers who reported teaching heavy loads were more inclined to talk about assessment negatively (Crusan, Gebril, & Plakans, 2016).

Nevertheless, it is still important to develop assessment literacy because the field of language testing is in constant state of flux (Malone, 2013). The advantages of developing assessment literacy is language educators will be able to identify appropriate assessment for specific purposes and analyze empirical data to improve their instruction (Coombe, Troudi, & Al-Hamly, 2012). Consequently, there is an emerging priority, to provide professional development for in-service and pre-service foreign language teachers so that they become literate in assessment analysis (Boyle, 2006). Furthermore, she asserted that foreign language professionals need a toolbox filled with skills and strategies that will enable them to decode assessment results, analyze their meaning, respond to what the results reveal, and apply them in teaching and evaluation program. Thus, to assess students' performance effectively, teachers have to know, understand, and apply the basic principles of language assessment in their every day practices in their classrooms, such as have the skills to choose appropriate assessment method, design valid assessment tasks, provide feedback to their students about their performance, and evaluate the process of teaching and learning which reflect the application of five principles of language assessment.

Considering the principles of language assessment, Brown (2004) states that the five principles of practicality, reliability, validity, authenticity, and washback provide useful guidelines for both evaluating an existing assessment procedure and designing one on teachers' own. In this study, the focus will be on what we traditionally think of as a 'test' rather than the broader notion of 'assessment' (Brown, 2001). Harmer (2007b) suggests that a good test is valid, should have marking reliability, practical, and have a marked washback/ backwash effect. Johnson (2001) asserts that a good test should possess validity, reliability, discrimination, and feasibility (same with practicality in Brown, 2004).

1. Practicality. Practicality is determined by the teacher's and students' time constrains, costs, and administrative details, and what occurs before and after the

test (Brown, 2004). The test will be worthless if it is so long that no one has the time to do it (Harmer, 2007b).

2. Reliability. A reliable test is consistent and dependable (Brown, 2001 & 2004) and it should be fairly easy to mark, but anyone marking it should come up with the same result as someone else (Harmer, 2007b).

3. Validity. (a) If the test measures what should be measured, it does what it says it will, and it can be seen as representative of the subject area being covered, then the test can claim content validity (Brown, 2001, Harmer, 2007b, & Johnson, 2001). (b) Face validity is the extent to which "students view the assessment as fair, relevant, and useful for improving learning" (Gronlund, 1998, p.210). Face validity is to do with what the teacher and students think of the test (Johnson, 2001). They should think it looks like the real thing (Harmer, 2007b). A term that is associated with face validity is "biased for best", means on how the students views the test to a degree of strategic involvement on the part of student and teacher in preparing for, setting up, and following up on the test itself (Brown, 2004).

4. Authenticity. Bachman and Palmer (1996, p.23) defined authenticity as the degree of correspondence of the characteristics of a given language test task to the features of a target language task, for example the task of the test should be related to "real world" task.

5. Washback. In large-scale assessment, Brown (2004) defined washback generally as the effects the tests have on instruction in terms of how students prepare for the test. He also stated that the challenge to teachers is to create classroom tests that serve as learning devices through which washback is achieved. The washback effect occurs when teachers see the form of the test their students are going to have to take and then, as a result, start teaching for the test (Harmer, 2007b).

Recently, Coombe, Troudy, & Al-Hamly (2012) suggested *The Seven Standards for Teacher Development in Assessment* developed by the American Federation of Teachers, the National Council on Measurement in Education and the National Education Association. Teacher should be skilled in choosing assessment methods appropriate for instructional decisions; developing appropriate assessment methods; managing, scoring, and interpreting the results of both externally produced and teacher-produced assessment methods; using assessment results when making decision about individual students, planning, teaching, developing curriculum, and involving students; developing valid grading procedures which use student assessment; communicating assessment results to students, parents, and other stakeholders; and recognizing unethical, illegal, and inappropriate assessment methods and uses of assessment information.

Between the principles and the standards, what is more important? Both of them are important because actually they have the similar purpose. For example, in standard 1, about choosing assessment methods appropriate for instructional decisions, it is the same with designing a valid test for students. Additionally, when we talk about assessment, it cannot be separated with teacher role as the one who develop, design, or evaluate assessment. Thus, it is necessary for teachers to understand and apply the principles of language assessment in evaluating the existing assessment procedure, or designing a new one by themselves.

Considering the issue, theories, and previous researches mentioned before, this study would like to find out about teachers' assessment literacy in Indonesia; in terms of the teachers' level of preparation and the application of the principles of language

assessment in Indonesian classrooms. The purpose of this study are: (1) to investigate the level of EFL teachers' preparation in assessing students' performance, (2) to investigate teachers' practices in applying the principles of language assessment in evaluating existing test and designing a new one, and (3) to find out about teachers' perception of the usefulness of the questionnaire that covers the principles of language assessment.

Since this research used a mixed method, a survey was designed to get the quantitative data of the level of EFL teachers' preparation in assessing students' performance and their practices in applying principles of language assessment in evaluating an existing test and/ or designing a new test using 7 items of questionnaire (adapted from *Classroom Assessment Literacy Inventory* (CALI) that adapted by Mertler (2003) from the similar instrument called *Teacher Assessment Literacy Questionnaire*) that focused on EFL teachers' background and 31 items of questionnaire focused on teachers' practices in evaluating an existing test and/ or designing a new test that covers the principles of language assessment that was developed from Brown's (2004, 31-37) questions in evaluating existing tests and designing a test for teachers with 4-point Likert scale (1 for strongly disagree, 2 for disagree, 3 for agree, and 4 for strongly agree). To get the qualitative data about teachers' perception of the usefulness of the questionnaire that covers the principles of language assessment, open-ended questions were added. Also, the questionnaire was consulted to an expert to be validated.

Forty three in-service teachers consisted of 30 female and 13 male responded voluntarily to the questionnaire through EFL teachers group in Facebook and WhatsApp using Google form. The participants in this study varied from EFL teachers who teach in elementary school, junior and high school, and university which located in Indonesia. The quantitative result of the survey was analyzed by means of descriptive statistics in terms of frequency distribution, percentage, mean, and standard deviation. Meanwhile, the qualitative result from the open-ended questions was summarized, categorized, and analyzed for frequency of mention.

By answering the proposed research questions, this study is expected to give some meaningful contribution in the application of principles of language assessment that can help teachers to evaluate an existing test or design a good test. Moreover, this study is expected to help them in improving their instruction in creating a good test and use the test to help their students to improve their performance. Also, this study is expected to be a reference for other researchers who want to conduct a study with similar topic.

Finding and Discussion

Table 1 shows that most of the participants are qualified with bachelor degree, are teaching in high schools, and have teaching experience around 1-5 years.

Table 1 Teachers' background information

Education level		Teaching experience		Level at which the teacher teach	
B.A., B.Ed.	76.7 %	1-5 years	62.8 %	Elementary	20.9 %
M.A., M.Ed.	11.6 %	6-10 years	32. 6 %	Junior high	30.2 %
Others (i.e.		11-15 years	4.7 %	Senior high	32.6 %
associate		More than 15		University	23.3 %
degree)	12. 6 %	years	-	Others (i.e.	
				kindergarten)	7 %

There are three major findings based on the purpose of the study. First is the teachers' level of preparation for assessing students' performance. Second is the teachers' practices in applying principles of language assessment. The last is the usefulness of the questionnaire that covers the principle of language assessment for teachers to evaluate and/ or design a test.

1. The level of preparation for assessing students' performance

From the teachers' level of preparation for assessing student performance that resulted from their undergraduate teacher preparation program, it was found that 79 % of them felt prepared (20.9 % of teachers felt very prepared and 58.1 % of teachers felt somewhat prepared) and 20.9 % of teachers felt somewhat unprepared. Moreover, to the best of their knowledge, 69.8 % of them took a course in language assessment as part of their undergraduate preparation and 30.2 % of them didn't take it.

2. Teacher's practices in applying principles of language assessment

Teachers' practices in applying practicality, reliability, content validity, face validity and 'biased for best', authenticity, and washback are presented in this part.

   a. Practicality

Table 2 below shows teachers' practices in applying practicality in evaluating and/ or designing a test. Most teachers agreed that they consider practicality in their practices (17.6 % strongly agree and 67.75 % agree, mean: 3.03 from 1-4 scale). Nevertheless, it can be seen from the result that there are some items showing teachers' disagreement in their practices but in lower percentage. The highest percentage of teachers' disagreement (30.20 %) is in administering the test smoothly without procedural 'glitches'. The second highest (11.60 %) are about students' capability in completing the test within the set time frame, the readiness of the material and equipment, determining methods for reporting results of the test in advance. Those items can be caused by teachers' level of preparation in designing a test. Preparing materials and equipment always needs time, and some teachers who are very busy sometimes prepare it in the last minute before the test, or even during the test. Moreover, some teachers tend to determine methods for reporting results of the test after right before they do the scoring. Yet, about students' capability in completing the test within the set time frame, I think there will be always one or two students in every class who cannot do that because of some reasons, like their proficiency level.

Table 2 Practicality

Item	Statement	SA	A	D	SD	M	Sd
				%			
Q1	I establish administrative details clearly before the test, such as the rubric of scoring and specification of the test.	30.20	62.80	7	-	3.2	.57
Q2	Students can complete the test I give, reasonably within the set time frame.	14	74.4	11.60	-	3.0	.51
Q3	I can administer the test smoothly without procedural "glitches".	14	55.80	30.20	-	2.8	.65
Q4	All materials and equipment of the test are ready, for example listening tape for listening section and answer sheets for students.	32.60	55.80	11.60	-	3.2	.64
Q5	The cost of the test is within budgeted limits.	20.90	65.10	14	-	3.1	.59
Q6	The scoring/ evaluation system of the test is feasible in the teacher's time frame.	7.10	83.30	7.10	2.4	3.0	.41
Q7	I determine methods for reporting results of the test in	4.70	83.70	11.60	-	2.9	.40

advance (before the test).							
Mean		17.64	67.75	13.30	2.4	3.03	.54

b. Reliability

Table 3 below shows teachers' practices in applying reliability in evaluating and/ or designing a test. Most teachers agreed that they consider reliability in their practices (34.42 % strongly agree and 57.22 % agree, mean: 3.28 from 1-4 scale). However, teachers' disagreement (the highest: 11.60 %) shows that they neglected some items in their practices, like making sure about the lighting, temperature, extraneous voice, and other classrooms conditions are equal and optimal for all students. This is a small thing, but may affect students' test result, like bad lighting may lead to students' misspelling, too cold temperature (because of the air conditioner) may make students feel uncomfortable and disrupt their concentration. Another item that had teachers' disagreement (11.60 %) was about objective scoring procedures that leave little debate about correctness of an answer. Sometimes, teachers found that their answer of a test might differ from their colleagues' answers. This item also needs to be considered because teacher has to make sure the key answers of the tests before distribute it to the students.

Table 3 Reliability

Item	Statement	SA	A	D	SD	M	Sd
			%				
Q8	I make sure that every student has a cleanly photocopied test sheet.	44.20	51.20	4.70	-	3.4	.58
Q9	I make sure that sound amplification is clearly audible to everyone in the room.	39.50	53.50	7	-	3.3	.61
Q10	I make sure that video input or power point presentation is equally visible to all.	41.90	53.50	4.70	-	3.4	.58
Q11	I make sure that lighting, temperature, extraneous voice, and other classroom conditions are equal (and optimal) for all students.	30.20	58.10	11.60	-	3.2	.63
Q12	Objective scoring procedures leave little debate about correctness of an answer.	16.30	69.80	11.60	2.30	3.1	.55
	Mean	34.42	57.22	7.92	2.3	3.28	.59

c. Content validity

Table 4 below shows teachers' practices in applying content validity in evaluating and/ or designing a test. Most teachers agreed that they considered content validity in their practices (17 % strongly agree and 71 % agree, mean: 3.05 from 1-4 scale). Brown (2004) states that the major source of validity in a classroom test is content validity. The two items below are two steps to evaluate content validity. The highest teachers' disagreement (16.3 %) is about the test specification. Although "it simply means that a test should have a structure that follows logically from the lesson or unit we are testing" (Brown, 2004, p.33), some teachers tended not to consider this item. I think it depends on teachers' creativity in designing a test because this item may offer students a variety of test items. The reason may lie in teachers' workload that prevents them from designing a test into number of sections.

Table 4 Content validity

Item	Statement	SA	A	D	SD	M	Sd
			%				
Q13	The classroom objectives are identified and appropriately	14	79.1	7	-	3.1	.46

253

	framed.						
Q14	The lesson objectives are represented in the form of test specification. For example, if you have to test students' all four skills in 30 minutes, so you specify how many minutes your students should spend for each skill.	20.9	62.8	16.3	-	3.0	.62
	Mean	17	71	12	-	3.05	.54

   d.   Face validity and 'biased for best'

Table 5 below shows teachers' practices in applying face validity and 'biased for best' in evaluating and/ or designing a test. Most teachers agreed that they considered face validity and 'biased for best' in their practices (16.11 % strongly agree and 70.48 % agree, mean: 2.98 from 1-4 scale). The interesting finding in this part is 37.3 % teachers' disagreement that the test has no 'surprises'. Thus, it means that these teachers tended to surprise their students in their test, but a teacher should have known that the test they give to his/ her students should covers the material he/ she has taught in the classroom. It can be strengthen by Brown (2004), who states that content validity is a very important ingredient in achieving face validity. That's why I think it is better if teachers test the students what they teach, not surprise them with may be unfamiliar materials. Brown (2001, p. 389) suggests that "if the test samples the actual content of what the learner has achieved or expects to achieve, the face validity will be perceived."

Table 5 Face validity and "biased for best"

Item	Statement	SA	A	D	SD	M	Sd
			%				
Q15	The directions of the test I make are clear.	25.6	74.4	-	-	3.3	.44
Q16	The ture of the test is organized logically.	17.3	61.5	3.8	17.3	2.5	.83
Q17	The test's difficulty level is appropriately pitched.	18.6	72.1	9.3	-	3.1	.53
Q18	The test has no "surprises".	4.7	58.1	32.6	4.7	2.7	.64
Q19	The timing of the test is appropriate.	20.9	74.4	4.7	-	3.2	.45
Q20	I offer students appropriate review and preparation for the best.	11.6	81.4	7	-	3.0	.43
Q21	I suggest strategies that will be beneficial for students.	20.9	72.1	7	-	3.1	.52
Q22	I structure the test so that the best students will be modestly challenged and the weaker students will not be overwhelmed.	9.3	69.8	20.9	-	2.9	.54
	Mean	16.11	70.48	12.19	11	2.98	.55

   e.   Authenticity

Table 6 below shows teachers' practices in applying authenticity in evaluating and/ or designing a test. Most teachers agreed that they considered authenticity in their practices (15.82 % strongly agree and 74.42 % agree, mean: 3.04 from 1-4 scale) even though there were some teachers that didn't apply some items in their classroom, such as the highest percentage in teachers' disagreement was providing some thematic organization in the test and interesting, enjoyable, and humorous test topics (16.3 %). From this result, it can be said some teachers might give a test to their students without considering how the language may be used in the real world, although most of the teachers consider it. Designing a test with considering authenticity will be more interesting for students. It's like linking the language with the real world so the test will not be boring and talking about unconnected items, such as reading passages from newspapers and listening from news report. Besides the students will be more

interested, they also will feel motivated and enthusiastic, or even challenged in completing the test.

Table 6 Authenticity

Item	Statement	SA	A	D	SD	M	Sd
			%				
Q23	The language in the test is as natural as possible.	25.6	72.1	2.3	-	3.2	.48
Q24	The test items are as contextualized as possible rather than isolated.	14	79.1	7	-	3.1	.46
Q25	The test topics and situations are interesting, enjoyable, and/ or humorous.	18.6	65.1	14	2.3	3.0	.65
Q26	Some thematic organization is provided in the test, such as through a story line or episode.	9.3	74.4	16.3	-	2.9	.51
Q27	Test tasks represent, or closely approximate, real world tasks.	11.6	81.4	7	-	3.0	.43
	Mean	15.82	74.42	9.32	2.3	3.04	.51

f. Washback

Table 7 below shows teachers' practices in applying washback in evaluating and/ or designing a test. Most teachers agreed that they enhanced washback in their practices (22.1 % strongly agree and 68.03 % agree, mean: 3.01 from 1-4 scale). Nonetheless, teachers' highest disagreement were in giving more than a number, grade, or phrase as their feedback when returning students' test (11.6 %) and gave students a chance to feedback on their feedback (11.6 %). These two points are not applicable in some cases, for instance, in the final examination or national examination. Because in Indonesia, after the final examination, teachers usually meet their students (or their parents) only when they give the report of the students' achievement in the whole semester or year, and it is usually done by their homeroom teachers so there will be no chance for the subject teacher to give specific comment or feedback.

Brown (2004) asserts that what happens before and after the test is critical, but many overworked (and underpaid) teachers return tests to students with only letter grade or numerical score and consider their job done. However, teachers need to give the appropriate feedback to the students that may help them to understand which part they are strong or weak to enhance or raise washback potential and it is good that most of the teachers do this in their practices. It also will be good if the final examination is not conducted in the very last meeting, so teachers will have chances to give feedback to the students.

Table 7 Washback

Item	Statement	SA	A	D	SD	M	Sd
			%				
Q28	I ask students to use the test results as a guide to setting goals for their future effort.	16.3	74.4	9.3	-	3.1	.51
Q29	I give comment generously and specifically on students' test performance.	30.2	62.8	7	-	3.2	.57
Q30	I give more than a number, grade, or phrase as my feedback when returning students' tests.	23.3	65.1	11.6	-	3.1	.59
Q31	I give students a chance to feedback on my feedback to seek clarification of any issues that are fuzzy, and to set new and appropriate goals for themselves in the future.	18.6	69.8	9.3	2.3	3.0	.62
	Mean	22.1	68.03	9.3	2.3	3.1	.57

3. The usefulness of the questionnaire

It was found that 27.9% of teachers thought that the 31 items questionnaire was very useful, 58.1% of teachers think that it was useful, and 14% of teachers think that it was less useful to evaluate or design a test for some reasons. Teachers' reasons are summarized in Table 8, 9, and 10.

The teachers who thought that the questionnaire they responded was very useful commented that it could help them to prepare and conduct a good test, and to evaluate a test to find the strength and the weaknesses of the test. Since assessment is one of the teacher's responsibilities, the questionnaire also could improve teachers' ability in teaching.

Table 8. Why teachers thought the questionnaire was very useful: comments summarized.

Usefulness	Teachers' comments
Very useful	*It can help us more as teachers to conduct a good test.*
	It can improve teacher's ability on teaching.
	It is very useful to know the quality (the strength and the weakness) of the test.
	It will guide us to evaluate or design a good assessment in the classroom.
	It's theoretical foundation for a teacher.
	It reminds me to do many things before, during, and after the test.
	It reminds me how I should prepare the test for my students.

The teachers who thought that the questionnaire they responded was useful stated that it was a kind of information and guideline to evaluate and/ or design/ created a good test and to increase their skill in teaching and learning process. These comments below are almost similar to teachers' comments that stated that the questionnaire was very useful.

Table 9. Why teachers thought the questionnaire was useful: comments summarized.

Usefulness	Teachers' comments
Useful	*It is to encourage English teachers to perform well in giving tests to students.*
	It can be guideline before setting a test and in conducting a test.
	It provides some information for us.
	It is useful because it can assess teacher's understanding in designing a test.
	It quite evaluates the way we test our students.
	It can help to increase the teachers' quality in teaching and learning process
	It reminds me of somewhat I've learned before.
	The items remind us as teachers to concern about certain things in assessing our students.
	It provides standard operating procedures to assess students learning.
	Teachers should learn to evaluate their students' performance in an appropriate method.
	It helps me much in designing a test.
	I may consider reminding my student about the upcoming test.
	It's useful because we can make the test organized to get the best result.

The teachers who thought that the questionnaire they responded was less useful mentioned that there were too many items to consider in evaluating and/ or designing a good test and sometimes the reality in the classroom is different and they tended to choose practical activity to evaluate and/ or design a test. Moreover, the items didn't cover the assessment method since they only covered the principle of language assessment.

Table 10. Why teachers thought the questionnaire is useful: comments summarized.

Usefulness	Teachers' comments
Less useful	*I think there are too many items to consider when a teacher is doing her test development and she should pay attention to all. Sometimes Indonesian teachers prefer to take practical steps.*
	It is too much.
	It's more useful for evaluating a test, but not for designing one.
	Sometimes, it doesn't reflect the reality.
	It has not covered assessment methods.

In addition, there were some difficulties that teachers face when they give assessment to students, especially in giving feedback. After the final test, teachers sometimes didn't give a feedback to the students. A teacher stated that feedback on the summative assessment was given as a class in general, because in certain institution students' test papers should be kept confidential by the institution, or can only be viewed during teachers and parents meeting. However, teachers can always give direct feedback during the teaching and learning process in the classroom.

T1: *It is about the feedback. After the final test, teachers sometimes didn't give a feedback to the students.*
T2*: Feedback on the summative assessment is given as a class in general, because students' test papers should be kept confidential by the institution, or can only be viewed during teachers and parents meeting. However, teachers can always give direct feedback during the class activities or formative assessment.*
T3*: Giving feedback to their result one by one is difficult because it's time consuming; But, I have tried it and found it worth it for them.*

It can be said that the questionnaire that the teachers responded somehow give them information about how to evaluate an existing test and/ or design a new one that can help them to produce a good test and improve their instruction to help students achieving the target language. However, I am sure that the activities written in these items are practiced or applied by the teachers whenever they evaluate or design a test. Some teachers might see that it was too many theories if they took a look at the questionnaire, yet, if the items are being practiced or applied well, it will not be too complicated but simple, does not need much time to do, and they are worth to do.

However, most of the previous study stated that teachers' assessment literacy was in the low level and did not feel adequately prepared in assessing their students from their undergraduate preparation program, for instance Wise et al. (1991) in Frey & Schmitt (2007), Mertler (2009), Diez (2010), Perry (2013), and Yamtim & Wongwanich (2014), and Lam (2015). However, those fndings were contradictory to the finding of this study.

The findings of this study showed that most of the teachers already had knowledge about assessment; and from their undergraduate program, they feel prepared in assessing students' performance. This finding was also supported by teachers' practices in applying the principles of language assessment in designing an existing test or creating a new one (total agree= 88.7 %, disagree = 13.9 %). In developing the language assessment literate teacher, Scarino (2013) stated that we need to take into account the processes through in which the teacher can critically examine their own assessment practices. Furthermore, Inbar-Lourie (2008) & Jeong (2013) stated that

teachers are expected to engage in classroom practices, report on learners' progress aligned with external criteria as well as prepare learners for external examination.

The possible reason to this contradictory finding can be caused by some reasons. First, it may lie in the item questionnaire used in the study. Most of the previous study used CALI (Classroom Assessment Language Inventory) as the questionnaire. When I took a look at the questionnaire, I think some of them are not appropriate for EFL teachers because there are items that cover about other subjects than English, like Math, Science, and Algebra. I think the items are more appropriate for classroom teachers like in elementary schools who teach all the subjects. Second, the courses in teachers' undergraduate program, like language testing and evaluation, EFL methodology, and language assessment, may change through the years and somehow they prepare teachers to be more prepared in assessing their students' performance. Crusan, Plakans, & Gebril (2016) stated that the reason of the difference finding might be due to changes over time in pre-service teacher education. Third, the development of the ICT or technology can also contribute to this finding. Teachers nowadays are easier to access many sources about assessment, like journals, books, or article provided on the internet and makes them to be more assessment literate. Harmer (2007a) suggests that newsletters and journals are a valuable way of keeping in touch with what is going on in the world of language teaching and some of them are published on the internet.

Additionally, despite of the usefulness of the questionnaire for most of the teachers, there were some teachers who think that the questionnaire was less useful because they thought there were too many items to be considered and the questionnaire didn't cover the assessment method. Therefore, assessment training from the government or educational board is still needed to improve teachers' assessment literacy, considering that (1) there are some items that play a role in producing a sound assessment which sometimes are being neglected by some teachers and (2) teachers' necessity to know and understand the appropriate method to produce sound assessment for the four skills of language: listening, speaking, reading, and writing.

Conclusion

Most of the teachers feel prepared in assessing their students and it was supported by their practices in applying the principles of language assessment in evaluating an existing test or designing a new test. The teachers' practices indicated that most teachers were literate in evaluating and designing a test since most of them applied the principles. However, regarding to the usefulness of the items provided in the questionnaire, most teachers thought that it was useful for them although there were some teachers thought that the items were too much and sometimes teachers didn't have enough time to consider all of those things.

From the findings, it is recommended for teachers to consider and apply the principles of language assessment in evaluating an existing test and/ or designing a new one so they will be able to produce a good test. Therefore, teachers will be able to improve their instruction and understand how to use the test to help their students in improving their success in the target language. Since this research is still general, about overall principles of language assessment, it is needed to have a further research about teachers' assessment literacy related to the method of assessing the four skills of language.

References

Bachman, L. F. & Palmer, A. S. (1996). *Language Testing in Practice*. New York: Oxford University Press.

Boyles, Peggy. (2006). Assessment Literacy. *New Visions in Actions: National Assessment Summit Papers*, 18-23.

Brindley, G. (2001). Language assessment and professional development. *Experimenting with uncertainty: Essays in honour of Alan Davies*, *11*, 137-143.

Brown, H. D. (2001). *Teaching by Principles: An Interactive Approach to Language Pedagogy*. Second Edition. New York: Pearson Education.

Brown, H. D. (2004). *Language Assessment: Principles and Classroom Practices*. New York: Pearson Education.

Coombe, C., Troudi, S., & Al-Hamly, M. (2012). Foreign and second language teacher assessment literacy: Issues, Challenges, and Recommendations. In C. Coombe, P. Davidson, B. O'Sullivan, S. Stoynoff (Ed.), *The Cambridge Guide to Second Language Assessment*: 20-29.

Crusan, D., Plakans, L., & Gebril, A. (2016). Writing Assessment Literacy: Surveying Second Language Teachers' Knowledge, Beliefs, and Practices. *Assessing Writing* , 28, 43-56.

DeLuca, C. & Bellara, A. (2013). The Current State of Assessment Education: Aligning Policy, Standards, and Teacher Education Curriculum. *Journal of Teacher Education*, 64 (4), 356-372.

Frey, B. B. & Schmitt, V. L. (2007). Coming to Terms with Classroom Assessment. *Journal of Advanced Academic*, 18 (3), 402- 423.

Gronlund, N. E. (1998). *Assessment of Student Achievement*. Sixth Edition. Boston: Allyn and Bacon.

Harmer, Jeremy. (2007a). *The Practice of English Language Teaching.* Malaysia: Pearson Education Limited.

Harmer, Jeremy. (2007b). *How to Teach English*. China: Pearson Education Limited.

Inbar-Lourie, Ofra. (2008). Constructing a Language Assessment Knowledge Base: A Focus on Language Assessment Courses. *Language Testing*, 25 (3), 385-402.

Jeong, H. (2013). Defining Assessment Literacy: Is It Different for Language Testers and Non-Language Testers?. *Language Testing*, 30 (3), 345-362.

Johnson, Keith. (2001). *An Introduction to Foreign Language Learning and Teaching.* England: Pearson Education Limited.

Lam, Ricky. (2015). Language Assessment Training in Hong Kong: Implication for Language Assessment Literacy. *Language Testing*, 32 (2), 169-197.

Malone, M. E. (2013). The Essentials of Assessment Literacy: Contrasts between Testers and Users. *Language Testing,* 30 (3), 329-344.

Mertler, C. A. (2003). *Classroom Assessment Literacy Inventory*. Retrieved October 11th, 2016, from http://pareonline.net/htm/v8n22/cali.htm

Mertler, C. A. (2003). Preservice Versus In-service Teachers' Assessment Literacy: Does Classroom Experience Make a Difference?. *Paper presented at the annual*

meeting of the Mid-Western Educational Research Association, Columbus , OH (Oct. 15-18).

Mertler, C. A. (2009). Teachers' Assessment Knowledge and Their Perception of the Impact of Classroom Assessment Professional Development. *Improving Schools*, 12 (2), 101-113.

Mertler, C. A., & Campbell, C. (2005). Measuring Teachers' Knowledge & Application of Classroom Assessment Concepts: Development of the *Online Submission*.

Newfields, Tim. (2006). Teacher Development and Assessment Literacy. In *Authentic Communication: Proceedings of the 5th Annual JALT Pan-SIG Conference* (48-73).

Perry, M. Lee. (2013). Teacher and Principal Assessment Literacy. *Theses, Dissertations, Professional Papers*. Paper 1391.

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental?. *Theory into practice*, *48*(1), 4-11.

Scarino, A. (2013). Language Assessment Literacy as Self-Awareness: Understanding the Role of Interpretation in Assessment and in Teacher Learning. *Language Testing*, 30 (3), 309-327.

Stiggins, R. (2014). Improve Assessment Literacy Outside Schools Too. *Phi Delta Kappan*, *92* (2), 67-72.

Stiggins, R. J. (1995). Assessment Literacy for the 21st Century. *Phi Delta Kappa*, 77 (3), 238-245.

Stiggins, R. J. (2001). The Principal's Leadership Role in Assessment." *NASSP bulletin* 85.621: 13-26.

Stiggins, R. J. (2002). Assessment Crisis: The Absence of Assessment for Learning. *Phi Delta Kappan*, *83*(10), 758-765.

Volante, L., & Fazio, X. (2007). Exploring Teacher Candidates' Assessment Literacy: Implications for Teacher Education Reform and Professional. *Canadian Journal of Education*, *30*(3), 749.

Webb, N. (2002). Assessment Literacy in a Standards-based Urban Education Setting. In *Annual Meeting of the American Educational Research Association, New Orleans*.

Yamtim, V. & Wongwanich, S. (2014). A study of classroom assessment literacy of primary school teachers. *Procedia – Social and Behavioral Sciences*, 116, 2998-3004.