JOLLT Journal of Languages and Language Teaching
https://e-journal.undikma.ac.id/index.php/jollt/index
Email: jollt@undikma.ac.id
DOI: https://doi.org/10.33394/jollt.v%vi%i.7481

April 2023. Vol.11, No.2
p-ISSN: 2338-0810
e-ISSN: 2621-1378
pp. 225-237

EVALUATING THE QUALITY OF A TEACHER'S MADE TEST AGAINST FIVE PRINCIPLES OF LANGUAGE ASSESSMENT

[1*]Dedi Sumarsono, [1]Moh. Arsyad Arrafii, [1]Imansyah

[1]English Language Education, Mandalika University of Education, Indonesia
*Corresponding Author Email: dedisumarsono@undikma.ac.id

Article Info

Article History
Received: March 2023
Revised: March 2023
Published: April 2023

Keywords
Classroom Assessment; Test; Principles of Language Assessment; Assessment Literacy

Abstract

Classroom assessment plays a dual role in both summative and formative functions, aiming to gather evidence, evaluate, and improve student learning. To ensure the accuracy and authenticity of assessment evidence, effective assessment instruments, such as tests, are essential for obtaining valid and reliable evidence of student learning. However, it is important to acknowledge that teachers may lack the necessary theoretical grounding to design and develop assessment instruments that align with sound language assessment principles. Consequently, this research study seeks to evaluate classroom-based assessment instruments, specifically language tests, against five fundamental principles of language assessment. Using an evaluative research approach, a teacher-developed assessment instrument was evaluated and rated against these principles. Data were collected through the analysis of language tests developed by teachers using documentation as a method. The study reveals that the teacher-made test generally meets all aspects of the principles, but some aspects require further attention. Accordingly, the study provides valuable insights and suggestions for improvement to address these areas of concern.

How to cite: Sumarsono, D., Arrafii, M.A., & Imansyah. (2023). Evaluating the Quality of a Teacher's Made Test against Five Principles of Language Assessment, *JOLLT Journal of Languages and Language Teaching*, 11(2), pp. 225-237. DOI: https://doi.org/10.33394/jollt.v%vi%i.7481

INTRODUCTION

Assessment is a crucial process involving the systematic gathering and interpretation of information concerning student performance, through a variety of methods and techniques. Its primary purpose is to provide reliable and relevant data that can inform both teachers and students. Teachers use assessment to make informed judgments about learners' progress, based on specific task criteria (Chapelle et al., 2015; Bajuti, 2018), and provide valuable feedback to enhance their teaching methods (Follmer & Sperling, 2019). Additionally, assessment enables teachers to determine appropriate next steps in the teaching and learning process. For students, assessment offers invaluable insights into their areas of strengths and weaknesses and provides guidance for achieving their learning goals through constructive feedback from their teachers (Harding et al., 2015; Su, 2020). In summary, assessment plays a pivotal role in enhancing the quality of teaching and learning, providing critical feedback to both teachers and students, and facilitating the achievement of educational objectives.

The effectiveness of assessment in fulfilling its intended functions hinges on the quality of the assessment instrument employed to collect information about student learning (Williams et al., 2022; Aprianoto & Haerazi, 2019). As such, the use of high-quality instruments is critical. An effective instrument is one that conforms to the standards for quality and the principles of classroom assessment. These principles include practicality, reliability, validity, authenticity, and washback (Brown and Abeywickrama, 2018). Given the unique nature of each classroom context, only classroom teachers possess the knowledge and skills necessary to design assessment tasks that align with the classroom characteristics, thereby grounding the task's development in the classroom context. Consequently, teachers'

assessment literacy and ability are paramount in the development of effective assessment instruments and practices (Stiggins, 1995, Popham, 2011, & Arrafii, 2021).

The quality of teachers' assessment relies heavily on teachers' understanding of assessment methods and formats, knowledge with regard to the test item construction, and understanding of classroom assessment principles underlying the best practice (Giraldo, 2018; Fulcher, 2012). However, little is known about teachers' ability in designing classroom test. In many cases, teachers normally use a test from publisher or textbook to measure their students' performance (Fulcher, 2012). It is rarely the case that they have developed assessment instrument themselves for their pedagogic use. If there is, the quality issue remains in question. This research aims to collect and evaluate teachers' made tests against five principles of language classroom assessment and indicate the levels of teachers' assessment literacy.

Principles of language classroom assessment

When designing a language assessment task, there are five fundamental principles that need to be considered by test developers including teachers to ensure that their products are able to achieve its purposes. The principles include practicality, reliability, validity, authenticity, and washback (Hughes, 2003; Brown & Abeywickrama, 2018).Each principle is described further below.

Practicality

When asking if a certain assessment is feasible, we want to determine whether it is possible, or practical, to use it in our current teaching situation. This principle is concerned with the "logistical, down-to-earth, administrative issues involved in making, giving, and scoring an assessment" (Brown & Abeywickrama, 2018, p. 26).Money, time, and resources at school can have a significant influence on the kinds of assessment teachers are able to use. Understanding the teaching context will therefore both guide and constrain the choices the teachers are able to make about assessment (Graves, 2000).

Reliability

Reliability refers to the extent to which assessment results are consistent. When using a reliable assessment "you can be confident that someone will get more or less the same score, whether they happen to take it on one particular day or on the next…[but]the score is quite likely to be considerably different, depending on the day on which it was taken" (Hughes, 2003, p. 3) if assessment is unreliable. Theprincipal idea about reliability is to ensure that students achieve their scores or results because of their abilities, and not due to other factors. The reliability of assessment is dependent upon several factors, including students, graders, the way assessment administered and the nature of assessment itself (Brown and Abeywickrama, 2018)

Student Related Reliability

Personal characteristics and students background may influence assessment results. A student's knowledge of particular subjects, cognitive style, gender and ethnic background may play a role in determining their results. There are also several temporary or random factors which may affect the reliability of assessment. Students may be ill, tired, anxious or simply having a bad day, and this can cause results to vary every time an assessment is given (Brown & Abeywickrama, 2018). Teachers also need to be aware of students' knowledge of and strategies for taking tests or other kinds of assessment. Students may be very familiar with certain types of assessment or may have had a significant amount of practice or preparation before taking them. Some students may also have developed effective strategies for completing assessments, such as predicting the correct answer for multiple choice test questions (Brown & Abeywickrama, 2018; Davies et al., 2002).

Rater Reliability

The rater is the person who marks, scores or judges an assessment. In many cases this person will be the teacher but in some cases the rater may be a professional language tester. As Davies et al. (2002) point out, raters are human and therefore capable of making mistakes which may influence assessment results. There are two aspects of rater reliability: inter-rater reliability and intra-rater reliability. The former refers to the similarity of scores given by two or more different raters to the same assessment. Differences in scores may be due to factors such as unfamiliarity with the criteria or scoring system, lack of attention to the criteria or scoring system, inexperience, fatigue or biases (Brown & Abeywickrama, 2018). The later refers to a single rater's consistency over a number of assessments (Brown & Abeywickrama, 2018). Once again, there are a number of factors which may lead a rater to either apply a different set of criteria to each assessment or to apply the same criteria differently. These may include fatigue, the sequence assessments are marked in, and bias towards students one may perceive as 'good' or 'bad'.

Assessment Administration Reliability

The conditions under which assessment occurs can also affect its reliability. Factors such as noise, lighting, legibility of test papers and condition of classroom furniture can lead to inconsistencies in assessment scores. Brown and Abeywickrama (2018) describe a situation in which noise coming from the street outside the classroom can prevent students hearing a tape recording during a listening comprehension test. The score these students received would more likely be a reflection of the interference of street noise rather than their listening comprehension ability.

Assessment Reliability

Finally, certain characteristics of the assessment itself can contribute to unreliability. Time limits, length of assessment, ambiguous questions and unclear instructions are among such factors (Brown & Abeywickrama, 2018). To take the example of time limits, students will take different amounts of time to complete tasks so if time runs out before a particular student is able to finish the assessment, his/her score will be affected. Similarly, time limits may influence the way students respond to the tasks. If for example a writing test requires students to write two essays in one hour, they might write the first very quickly, so they can complete the second essay.

Validity

Validity refers the credibility and trustworthiness. In the context of assessment, a valid assessment measures what it aims to measure, "does what it is intended to do" (Davies et al. 2002, p. 221). Another aspect of validity is the interpretations or uses of assessment results. Determining assessment being valid or not is not an easy task (Brown, 2004). However, people can look at several sources of evidence to help make the decision about the validity of a test. These sources of evidence can be in the formscontent validity, criterion validity, construct validity, consequential validity, and face validity (Brown and Abeywickrama, 2018).

Content Validity

Content validity relates to the content of an assessment. In short, the content of an assessment – questions, tasks and subject matter – should reflect the ability teachers are trying to assess (Brown, 2004). Hughes (2003) offers an example: it is obvious that a grammar test, for instance, must be made up of items relating to the knowledge or control of grammar. But, this in itself does not ensure content validity. The test would have content validity only if it included a proper sample of the relevant structures. Just what are the relevant structures will

depend, of course, upon the purpose of the test. It is less likely that an achievement test for intermediate learners to contain just the same set of structures as one for advanced learners. If on the other hand a teacher was "trying to assess a person's ability to speak a second language in a conversational setting, asking the learner to answer paper-and- pencil multiple-choice questions requiring grammatical judgements does not achieve content validity" (Brown & Abeywickrama, 2018). Therefore, if the content of an assessment matches the ability it is supposed to assess, then it has content validity (Brown, 2004).

Criterion Validity

Criterion validity refers to the relationship between assessment results and other indicators of language ability. In this case, if the results of the assessment that teachers use coincide with some other criterion, or benchmark, which are believed to provide a good indication of language ability, the task has criterion validity. There are two aspects of criterion validity: concurrent validity and predictive validity. Concurrent validity refers to how a student's performance on a particular assessment compares to his/her performance on other measures of language ability at roughly the same time as the assessment was taken. The achievement of similar results on different assessments can also demonstrate concurrent validity. For example, a student may receive a high score on a classroom listening comprehension test and shortly afterwards receive a high score on the listening component of the IELTS exam. Predictive validity is the extent to which an assessment can predict how well an individual will be able to perform a particular task in the future.

Construct Validity

Language ability or proficiency is not a directly accessible tangible trait, and that language assessment is based on one's view on the nature of language ability. A construct is a concept or definition of language ability and therefore concerned with how well an assessment represents the concept or definition of language ability upon which it is based (Bachman, 1990). The test developer must spell out just what that construct is or what it consists of. The test can be valid only if the test construct is a complete and accurate picture of the skill or ability it is supposed to measure. For test to have construct validity, the tasks a student is required to perform must be consistent with our definition of language ability. In other words, an assessment must "tap into" our concept of language ability (Brown, 2004, p.25).

Consequential Validity

It is important to bear in mind that language assessment does not occur in isolation but it is used within a broader social context and is used on people. This means teachers have to consider the consequences of language assessment (consequential validity). Some issues need to be considered regarding consequential validity includes whether the evidence of assessment works well enough to make appropriate decisions regarding students' learning, the types of language abilities valued or perceived as important in assessment, the extent to which assessment results can be used as a reference to judge the potential performance of students in real life, the potential impacts of assessment on classroom instruction. All of these factors are likely to influence our decision about whether or not to use a certain type of assessment (Bachman, 1990).

Face Validity

Face validity is a critical concept in the realm of assessment, which pertains to the extent to which an assessment appears to measure what it claims to measure based on its physical characteristics. For instance, a reading comprehension assessment that entails reading a short newspaper article and answering questions about it is likely to appear to

measure reading comprehension (Bachman, 1990). Face validity is determined by the perceptions of teachers, assessors, and students, and if an assessment item appears to be well-designed, it is deemed to possess face validity (Heaton, 2000). However, it is important to note that face validity alone is inadequate to establish the validity of an assessment. Nonetheless, it is still crucial because the perceptions of stakeholders will affect their reactions to the assessment (Bachman, 1990). An assessment without face validity may result in numerous challenges, as Hughes (2003) elucidates that such a test may not be accepted by candidates, educators, education authorities, or employers, and if it is utilized, the candidates' response to it may not accurately reflect their abilities.

Authenticity

An important factor that requires careful attention is the authenticity of the language and tasks used in language assessments. The degree to which the language and tasks used in an assessment are representative of real-life situations is referred to as "authenticity." In other words, educators must analyze whether the evaluation accurately captures actual language use in natural settings. The definition of authentic language is "oral or written language examples that are not consciously produced for instructional reasons" (Nunan, 1999). Likewise, genuine activities are those that ask students to act in a way that closely mimics how they behave in actual life situations. For a precise evaluation of language competency, real language and tasks must be included in language assessments.

Washback

When choosing or designing assessment, teachers need to ask whether the assessment will create a positive influence on their teaching and students' learning experience. As assessment can exert a powerful influence on teachers, learners and society in general,it is not surprising therefore that assessment can affect the nature of English language teaching and learning, including what aspects of the language are taught, the amount of time spent on particular aspects of the language and the teaching methodology used in classrooms. The effect an assessment has on teaching and learning is known as washback or backwash. An assessment can have either a positive or negative effect (McNamara, 2000).

These five principles, when they were considered during the assessment design and development, can ensure effective classroom assessment. As the classroom teacher is a member of classroom community who interacts and engages with the classroom discourse intensively, teachers are considered the most knowledgeable person about the classroom context and thus are likely to become the best assessor of student learning. Due to a prolonged engagement in the classroom activities, compared to the outsiders, teachers can gauge more accurate and comprehensive evidence of students' learning and development. However, until to date, we have a limited understanding about teachers' performance in designing and developing classroom assessment instrument which helps them make the right decision regarding students' learning. What we have known is that rather than developing their own instrument, teachers frequently adopt available instrument from the textbook to measure their students' learning progress and achievement (Fulcher, 2012). Many of such adoption were proceeded without a proper adaptation and modification.

Additionally, although some teachers reported that they have developed their own assessment instrument to capture students learning, the quality issue regarding this instrument remained persist (Fulcher, 2012; Popham, 2011). We still do not get informed about the extent to which these self-made assessment instrument addresses the quality issues and principles of an effective instrument. This research was framed and guided by the following research question: To what extend have the principles of language assessment been incorporated in the teacher's made tests?

RESEARCH METHOD

Given the purpose to uncover the quality of teachers' made assessment against five principles of language classroom assessment, this research employed a case study design (Yin, 2009) by which one of the assessment tasks from a teacher working at English department at higher education level was selected to be evaluated thoroughly. Thus, the unit of analysis however in this study was the teacher's made test instead of the teacher himself. The test was examined against the principles of language classroom assessment (Brown and Abeywickrama, 2018).

Data Gathering Method

The study gathered teachers' assessment instruments which have been used by the teachersto gather evidence of student learning. In this study, assessment instrument refers to the teachers' made language test for use in summative formal assessment such as mid or final semester assessment. This criterion excludes daily assessment tasks and exercises that were used to monitor student learning. A set of English tests was gathered from the participating teachersfrom English education department, Universitas Pendidikan Mandalika, Lombok. Researchers contacted the teachers in person and asked a copy of their assessment tasks. The researchers ensure that the assessment artifacts collected are teacher-made tests to ensure the trustworthiness and accuracy of the research findings. If the artifact is proven to be taken from a publisher, it was excluded for analysis. Given this inclusion criteria, a number of teachers' made tests were brought forward for analysis. However, in this report we present the results of our evaluation on a single assessment task which was an essay writing task for sophomore students at English department (see appendix 1 for the details of the assessment task). This task was selected due to its features that characterise an effective assessment instrument and maybe considered as a model of good assessment task of essay writing.

Data Analysis Method

To evaluate the quality of teachers' assessment instruments, data from this research were analysed using qualitative and quantitative methods. Qualitative method is a method used to describe data using words but may also using scores (quantitative) to develop a robust analysis to support the qualitative description of the data. The usage of both methods is considered powerful to strengthen the description of the research data. To do this, comparative and contrasting analysis methodswere used (Mahsun, 2017). Initially, the assessment task was read, annotated and evaluated against the principles of assessment design and development (Brown and Abeywickrama, 2018). This process requires a description sheet to capture evidence of congruence between teachers' assessment instruments and the principles. To indicate the quality of assessment instruments, the instrument was rated in a scale from 1 (poor) to 5 (excellent) (see appendix 2for the details of this rating system). The teachers' made assessments, which have been rated, were tabulated and computed to measure the overall quality of the assessment instrument. From this evidence, a category of the quality of assessment instrument was developed.

To ensure trustworthiness of the analysis, intra-rater and interrater strategies wereemployed. Initially, the principal investigator analysed the data and then revisited them a few weeks later. Then, the same data was evaluated by the co-researcher. Then we sit together to discuss the data to arrive at final evaluation. The results of our final, agreed analysis (scoring) can be seen in appendix 2.

RESEARCH FINDINGS AND DISCUSSION

To answer the research question regarding the extent to which language assessment principles were incorporated in the teachers' classroom assessment instrument, teachers' assessment artifactswere collected and analysed according to five language assessment principles proposed by Brown and Abeywickrama (2018) that include practicality, reliability,

validity, authenticity, and washback. The teacher made test in this study is described and connected in relation to each principle of language assessment. The results of analysis were displayed in the appendix 2:

Practicality

When dealing with the issues of educational resources within and outside the classroom, issues regarding practicality principle of this assessment can be considered high(rated 5 out of 5) because teachers do not need to spend a large amount of money to prepare and apply the test in the classroom. Neither do they need equipment for designing, collecting, administering, and evaluating ofstudents' work from this test. The test just requires several pieces of paper.Teacher could also administer the task either orally through dictation of the instructions and question or written through providing students with the writing instruction and questions on the board. For this reason, this test can be used in different educational contexts, e.g., remote or urban schools, or school with either high or low socioeconomic status, small or big classroom size.

In terms of time spent for designing the test, it requires relatively short time. However, designing assessment rubric that mirrors students' ability in writing expository essay might be challenging and time consuming for some teachers. In addition, time spending on markingstudents' work based on this test might be another issue associated to practicality principal. Teachers can use holistic scoring rubric which believed to be effective rubric of students' writing ability (Brown & Abeywickrama, 2018). However, teachers need more time to grade the task, especially when it is employed in a large classroom size. In this regards, this test is therefore less practical compared to other types of assessment such as multiple choices. In this aspect the practicality of this test is rated 3 out of 5)

Reliability

The overall score for the reliability principle of assessment is 3.6/5.0. In terms of student-related reliability, this test reliability can be considered high(score 4/5) because it contains the topic which is closely related to their life and past experience.It is assumed that students may have had background knowledge of the topic of presented in test items. The test asks students to recall their childhood experience related to the most favourite game they played as a kid. This kind of live experience is unique for each student. This will encourage and motivate them to write because are knowledgeable about the topic.

With regards to inter-rater reliability, this reliability issue of this assessment was score 5 out of 5 because it is dependent upon the raters' quality, experience, perspectives and perhaps qualifications which can influence they students' work graded(Weigle, Boldt, and Valsecchi,2003). Inter-rater reliability could be low if the raters have had different perspectives in assessing students' work and used different contents of rating scales. On the other hand, inter-rater reliability can be high if two or more raters have had an agreement for assessment criteria to be used (Hughes, 2003).

Regarding assessment administration reliability, assessment reliability relies heavily on several factors such as noise, classroom size, number of students taking the test. When this test is delivered in a non-disrupted atmosphere, its administration reliability can be high, or otherwise. Therefore, we score this aspect 5/5. Assessment reliability of this test can be high because it has a clear understandable instruction, followed by several scaffolding questions which limit students' freedom in providing response (Hughes, 2003). However, a further instruction telling students about this should be displayed so students know more what to include in their essay.For this reason, assessment reliability was scored 4/5.

Validity

We rated 4 out of 5 for the validity of this assessment overall. All validity aspects were score 4 each. With regard to the content validity,thevalidity of this testcan be high

because the questions, tasks and subject matter reflect the extensive writing skills to betested, indicated by a number of words required (300 words). In addition, there is a match between the targeted writing genre and instruction as well as questions being presented in this assessment. The task for students is to write long stretch of an expository essay, complemented with a sequence of instructions that guide students to compose an extensive writing. However, with regards to the criterion validity, the validity of these assessment might be considered low due to the absence of other measures of language abilities. Nevertheless, the predictive validity might be achieved as the result of this assessment can be a reference to predict students' ability in writing an extensive textin the future occasion.

In terms of construct validity, this test is valid because the taskis consistent with the definition of language ability targeted in some ways: *Firstly,*extensive writing is a long scratch of writing that is well-connected, and this assessment asks students to write a 300-word expository essay which meets the criteria of extensive writing. *Secondly,* expository essay asks students to clarify information and provide reasons to the readers. This test asks the test takers to describe and explain why a particular game in their childhood had become a favourite game they played when they were kid. *Thirdly*, the task provided students some assessment rubric and this leads students to know the structure of the essay being assessed. If students know these assessment components, they can focus on them when developing their essays and therefore it improves the construct validity of the assessment.

The test analysed here may be considered to meet sequential validity because it can be used to make judgment and decision about students' performance. With comments and grades from markers or raters, it informs students how well they learn. Moreover, the assessment results can be used by teachers as a reference to improve their instruction, serving a formative function of assessment (Black and Wiliam, 2018). Dealing with face validity, this test meets the validity principles. its instruction and question require students to write, not to listen nor reading nor speaking. There is also a clear writing genre to be composed by students. Hughes (2003) argues that good test is that the one tests only targeted skills, in this case writing ability. Similarly, Heaton (2000)asserts that a test can have face validity when teacher, students, and assessors feel alright to assessment items. As evaluator of this task, we agreed that the task preserves face validity.

Authenticity

Authenticity principle of language assessment can be evaluated from the extent to which the assessment languages is natural and clearly reflect the authentic real-life situation. For this principle, all aspects of authenticity principle were scored 5. We rated the test 5 overall for some reasons. First, it explicit from the test that the test items are presented in natural and communicative way. This leads to situation where the test instructions are understood well by test-takers and evaluators. Second, the test is contextualised as a context is presented in the topic or material of the test (Firman et al., 2021). Instruction of the test introduces the context of assessment in which the test-takers must relate the topic to their past (childhood) experience. Third, the topic is meaningful, interesting, and relevant to students because the topic is well-known by students and is closely related to them. O'Malley & Pierce (1996) found that that providing students an interesting, relevant and well-known by studentscan help students to improve their writing performance. In addition, writing a topic that is familiar with personal circumstance could increase students' interest (Pajares & Schunk, 2005)

As there is only one task to be performed by students, one might argue that the task is not presented thematically. However, it is not significant issue affecting the authenticity principle because there is clear sequence of instruction that can help students tocomplete the task. Moreover, the topic and the context are introduced early in instruction allowing students to activate their prior knowledge about the topic. This was followed by next instruction that leads to the focus of task. Then students are given follow-up instructions that provide students

an approach to finish the task, and finally are given reasonable amount of time to respond to the task.

Washback

A test can produce meaningful washback if it positively affects students and teachers about how well they do their jobs, determines what and how teachers teach and students learn, provides adequate time for preparation, gives learners constructive feedback that boosts language development, is more formative than summative oriented, equips opportunities for learners to achieve peak performance (Brown & Abeywickrama, 2018). Considering these features of positive washback, we gave 4.5 for the washback aspect of principle for this task.

Firstly, the task analysed in this study assesses pupils' capacity in composing an expository essay. This assessment tests students' ability in describing, explaining, and providing information to the readers. This target is reflected in the tasks instruction and questions that ask students to describe one of the most favourite games they played when they were kid and provides reasons for selection. Hence, the instruction and questions of the task meet particular characteristics of expository essay. It does not ask students to write other writing genres. Secondly, this assessment promotes opportunities for students to perform self- or peer assessment on their works (Dolba et al., 2022). This practice is suggested as it helps students to discover their learning weaknesses and strengths, and this subsequently improves performance (Baars et al., 2014; Haerazi & Kazemian, 2021). Research has indicated that training on the use of peer or self-assessment can be effective for improving their writing performance in the future task (Kostons et al., 2012). Further, this assessment provides written formative feedback for students which can be used by students to enhance their writing development.

Thirdly, as the test provides guided questions for students to complete the task, teachers and students understand the instruction and task are understood in the same way. In formative assessment practices, similar understanding between the teachers and students of instruction and learning target can increase learning opportunities and outcomes (Arrafii, 2021). Lastly, the test has provided some opportunities for student to practice and apply skills and knowledge in writing an expository essay. This competence is undoubtedly crucial and useful for their future life, especially for those who want to pursue a career as a journalist or writer. However, some characteristics of positive washback proposed by Hughes (2003) are not fully accommodated in this test. For example, this test merely provides single task that is to write expository essay, while lacking in providing multitasks.It is likely that if the task is single, students' preparation for the test might be reduced to include this type of ability or performance.

CONCLUSION

The teachers' made assessment reported in this paper can be considered as a very good example of assessment of writing task because it satisfies all five principles of classroom assessment although some minor issues still appear. However, the finding reported here cannot be applicable to other teachers as they may have developed different level of expertise with regard to assessment design and development. Therefore, as Popham (2011) has pointed out that only a small number of teachers have had adequate assessment literacy and the respondent of this research being considered belong to this minority, an interventionist research approach through professional development involving a large number of teachers is considered an effective way to improve teachers' assessment literacy, especially on area of test item design and development.

REFERENCES

Aprianoto, & Haerazi. (2019). Development and Assessment of an Interculture-based Instrument Model in the Teaching of Speaking Skills. *Universal Journal of Educational Research, 7*(12), 2796–2805. https://doi.org/10.13189/ujer.2019.071230

Arrafii, M. A. (2021). "We must assess all, even a student farting [is also assessed] for the behavioural aspect of learning": Teachers' conceptions of assessment in the context of assessment reform in Indonesia. *The Curriculum Journal*, 00, 1–23. https://doi.org/10.1002/curj.130

Arrafii, M. A. (2021) Assessment reform in Indonesia: contextual barriers and opportunities for implementation, *Asia Pacific Journal of Education*, DOI: 10.1080/02188791.2021.1898931

Baars, M., Vink, S., van Gog, T., de Bruin, A. &Paas, F. (2014) Effects of training self-assessment and using assessment standards on retrospective and prospective monitoring of problem solving, *Learning and Instruction*, 33, 92-107

Bachman, L. F. (1990) *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Baiutti, M. (2018). Fostering assessment of student mobility in secondary schools: Indicators of intercultural competence. *Intercultural Education*, *29*(5–6), 549–570. https://doi.org/10.1080/14675986.2018.1495318

Black, P. & Wiliam, D. (2018) Classroom assessment and pedagogy, *Assessment in Education: Principles, Policy & Practice*, 25:6, 551-575

Brown, H. D. & Abeywickrama, P. (2018) *Language assessment: Principles and classroom practices* (3rd ed.). New York: Pearson Education Inc

Brown, H. D. (2004) *Language assessment: Principles and classroom practices*. New York: Pearson Education Inc.

Chapelle, C. A., Cotos, E., & Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing*, *32*(3), 385–405. https://doi.org/10.1177/0265532214565386

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (2002) *Dictionary of language testing*. Beijing: Foreign Language Teaching and Research Press.

Denscombe, M. (2007) *The good research guide: For small-scale social research projects* (3rd edition) Buckingham, UK: Open University Press.

Dolba, S., Gula, L., & Nunez, J. (2022). Reading Teachers: Reading Strategies Employed in Teaching Reading in Grade School. *Journal of Language and Literature Studies*, *2*(2), 62–74. https://doi.org/10.36312/jolls.v2i2.874

Firman, E., Haerazi, H., & Dehghani, S. (2021). Students' Abilities and Difficulties in Comprehending English Reading Texts at Secondary Schools; An Effect of Phonemic Awareness. *Journal of Language and Literature Studies*, *1*(2), 57–65. https://doi.org/10.36312/jolls.v1i2.613

Follmer, D. J., & Sperling, R. A. (2019). Examining the Role of Self-Regulated Learning Microanalysis in the Assessment of Learners' Regulation. *The Journal of Experimental Education*, *87*(2), 269–287. https://doi.org/10.1080/00220973.2017.1409184

Fulcher, G. (2012) Assessment Literacy for the Language Classroom, *Language Assessment Quarterly*, 9, 2, 113-132,

Giraldo, F. (2018) Language assessment literacy: Implications for language teachers. *Profile: Issues in Teachers' Professional Development,* 20, 1, 179-195

Graves, K. (2000) *Designing language courses: A guide for teachers*. Boston: Heinle and Heinle.

Haerazi, H., & Kazemian, M. (2021). Self-Regulated Writing Strategy as a Moderator of Metacognitive Control in Improving Prospective Teachers' Writing Skills. *Journal of Language and Literature Studies*, *1*(1), 1–14. https://doi.org/10.36312/jolls.v1i1.498

Harding, L., Alderson, J. C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing*, *32*(3), 317–336. https://doi.org/10.1177/0265532214564505

Heaton, J. B. (2000) *Writing English language tests* (new ed.). Beijing: Foreign Language Teaching and Research Press.

Hughes, A. (2003) *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.

Kostons, D., van Gog, T. &Paas, F. (2012) Training self-assessment and task-selection skills: A cognitive approach to improving self-regulated learning, *Learning and Instruction* 22, 121-132

Mahsun. (2017). *Metode penelitianbahasa.*Depok: PT Rajawali Pers.

McNamara, T. (2000) *Language testing*. Oxford: Oxford University Press.

Nunan, D. (1999) *Second language teaching and learning*. Boston: Heinle and Heinle.

Popham, W. J. (2011) Assessment literacy overlooked: A teacher educator's confession, *The Teacher Educator*, 46, 4, 265-273.

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

Su, H. (2020). Educational Assessment of the Post-Pandemic Age: Chinese Experiences and Trends Based on Large-Scale Online Learning. *Educational Measurement: Issues and Practice*, *39*(3), 37–40. https://doi.org/10.1111/emip.12369

Williams, T., Wiener, J., Lennox, C., & Kokai, M. (2022). Lessons Learned: Achieving Consensus About Learning Disability Assessment and Diagnosis. *Canadian Journal of School Psychology*, *37*(3), 215–236. https://doi.org/10.1177/08295735221089457

Yin, R. K. (2009) *Case Study Research: Design and Methods* (4 ed.). London: SAGE Publications, Inc.

Appendix 1:A sample of teacher-made test designed to measure students' performance in writing expository essay

Assessment Questions	Everyone had favourite game to play when they were kid. What was your favourite game when you were kid?
Instructions	The following is one of approaches you may use to complete this task. a. Before you begin to write your essay, think about one of the most favourite games you played when you were kid. (5 minutes). b. Describe that game: What? How? When? Where? Why? (10 minutes) c. Write a three-hundred-word essay that tells about one of the most favourite games you played when you were kid. Provide the reason why you like it most and support your ideas with specific examples and details (40 minutes). d. You will have 70 minutes to complete your essay. e. Revise and review your essay within last 15 minutes of your time. f. Your essay will be assessed based on the content, organisation, rhetorical discourse, grammar/mechanics, and word count.
Materials	Students are given a piece of paper that contains instruction and assessment questions. In this paper, space for written response is also provided.
Marking/Scoring/Feedback	To mark this assignment, holistic scoring, which ranges from 5-0, would be used. The descriptors of each score are presented below: Essay with score of 5 is categorized as an excellent essay because a high degree of proficiency in response to the assignment is presented. However, a few errors/slips might be found. Essay with score of 4 is classified as a very good essay because a clear proficiency in response to the assignment is presented, but it may contain minor errors. Essay with score of 3 is considered as a competent essay because it demonstrates proficiency in response to the assignment. Essay with score of 2 is a limited essay because it indicates some degree of proficiency in response to the assignment, but it sometimes contains errors. Essay with score of 1 is categorized as a poor essay because it shows poor proficiency in response to the assignment. Essay with score of 0 is categorized as a failure essay because it reveals fundamental errors in writing skill. Detail criteria of each score are presented in the column below. Feedback will be given in written form on each element being rated (content, organization, rhetoric discourse, and grammar), and presented in the bottom of the essay. Feedback also points out the strengths and the weaknesses of the essay as well as offers some suggestions for strategies for improving students' writing. Besides that, marginal comments which address the issue of coherence, unity, supports, and examples of the essay, will be provided. In-text comments deal with grammatical errors. This will be done by pointing out some errors in text and suggesting the correct forms.

Appendix 2: The result of researchers' evaluation on teacher-made test (appendix 1) in each principle of language assessment

No	Criteria (Assessment principles)	Rating				
		1	2	3	4	5
1	**Practicality**					
	a) time for design, administration, marking			**X**		
	b) money					**X**
	c) resources/equipment					**X**
2	**Reliability**					
	a) student-related reliability				**X**	
	b) rater reliability (intra- and inter-rater)					**X**
	c) assessment administration Reliability					**X**
	d) assessment reliability				**X**	
3	**Validity**					
	a) content validity				**X**	
	b) criterion validity				**X**	
	c) construct validity				**X**	
	d) consequential validity				**X**	
	e) face validity				**X**	
4	**Authenticity**					**X**
	a) language is as natural as possible					**X**
	b) questions/tasks contextualised, not isolated					**X**
	c) topics meaningful, interesting, relevant to students					**X**
	d) questions/tasks organised thematically					**X**
	e) questions/tasks closely reflect real life					**X**
5	**Positive Washback**					
	a) assess abilities we want students to develop					**X**
	b) include wide range of questions/tasks					**X**
	c) vary questions/tasks over time				**X**	
	d) direct assessment					**X**
	e) criterion-referenced assessment					**X**
	f) assessment based on objectives				**X**	
	g) assessment well understood by students and teachers					**X**
	Total					

Note: Excellent (5); Very Good (4); Good (3); Satisfactory (2); Poor (1)