Jurnal Ilmiah Profesi Pendidikan
Volume 10, Nomor 4b, Edisi Khusus 2025
ISSN (Print): 2502-7069; ISSN (*Online*): 2620-8326

Addressing the Validity of a Teacher Made-Test Constructed by High School English MGMP Teachers in West Lombok

Bahrul Hadi*, Untung Waluyo, Henny Soepriyanti
Program Studi Pendidikan Bahasa Inggris, Jurusan Pendidikan Bahasa dan Seni, FKIP, Universitas Mataram, Jl. Majapahit No. 62, Mataram NTB, 83125. Indonesia
*Corresponding Author: hadibahrul195@gmail.com

Article History
Received : September 16th, 2025
Revised : October 23th, 2025
Accepted : November 20th, 2025

Abstract: In second language learning, assessment plays a vital role in evaluating and measuring students' proficiency levels based on the material taught. Therefore, this research aims to examine the strategies, practices, and challenges faced by an English MGMP teacher in West Lombok face during the development of a valid teacher-made test for Senior High School students. This study employs a qualitative design, and the data were collected through triangulation techniques such as teacher interviews and document analysis of the test item from two schools in West Lombok, six teachers from SMAN 1 Narmada, and four teachers from SMAN 1 Gerung. The findings indicate that the majority of teachers maintain the content validity by utilizing the curriculum, module, and test blueprint as their core references before and during the test development process. However, several test items were found to be misaligned with the curriculum and the materials taught. Furthermore, none of the respondents considered the alignment between the test with the specific language skills that being assessed. In addition, the majority of the problem that teacher usually face is the wide variation in students' proficiency levels, which makes it difficult to design a fair test for all students. The study concludes that, although the most of English teachers had ensured the content validity of their test items by referring to official documents such as the curriculum, teaching modules, and test blueprints during the test development process, there are several items which still deviated from the material taught, and none of teachers addressed whether the test formats appropriately matched the language skills being assessed such as speaking, writing, listening and reading. Their strategies mainly involved adjusting items to students' proficiency levels, collaborating with colleagues, and managing test banks through digital platforms such as WhatsApp and the PIJAR application. However, none reported using analytical measures like item discrimination index to examine test quality. The primary obstacle they faced was the broad variation in students' abilities, alongside many learners performing below the expected standard. This challenge was intensified by the absence of recent professional development initiatives or coordination from the English MGMP in West Lombok over the last one to two years.

Keywords: Content Validity, Cognitive Validity, Face Validity, Language Assessment, Teacher-made Test

INTRODUCTION

One of the problem was that the outcome of second or foreign language learning is generally not satisfactory and sometimes even disappointing, and this a matter which has caused much concern for both trainers and those responsible. Despite the fact that students are being taught English for many years, a majority of learners still express their disappointment with what they learn in the classroom. This situation remains in place due in large part to the lack of effective teaching and one major reflection of this poor quality if the way teachers conduct assessment. Thus, in L2 instruction, it is nature for teachers to construct tests which are capable of assessing students' true learning achievements with respect to the four vital language skills- listening, speaking, reading and writing. Additionally, assessments

are significant since they serve to measure the degree of students' attainment of instructional objectives. Given this, there are two general categories of tests – standardized tests and teacher-made tests – for different types of evaluation. Nevertheless, most frequently used are tests constructed by teachers and due to their feasibility and congruency with classroom performance. The results of these tests carry weight in grading, promotion and thus students' academic development and therefore the accuracy is all the more important. The quality of teacher-made tests also influences how well students' true abilities are represented fairly and consistently. In this sense, the debate on validity in teacher-made tests acquires great interest in teaching practices.

Messick (1989) stated that a test is valid if it measures what it purports to measure, and this encapsulated the essence of meaningful testing. Secondly, testing must be valid in that it should not only sample test what has been taught but also reflect the language skills to be tested. Put simply, one should measure a reading skill with a reading task and a speaking skill with speaking items that correspond to the learning objectives. Hence, a teacher will adopt specific procedures in designing an assessment such as planning of the test, item writing techniques and ensure that items measure what students are expected to learn. Moreover, teachers should field-test the test items for reliability and discover mistakes which can influence students' scores. It's also important that teachers do a thorough job of testing and grade objectively so the results are consistent. These steps allow teachers to assess the quality of each item and revise it if necessary. Thereby, a reliable and valid test can help improve the overall trustworthiness of learning evaluation. Therefore, understanding validity theories is not just a theoretical necessity; it is a practical one for everyday classroom assessment.

It is very important in engaging students toward developing their language skill, especially to enhance meaningful learning when they are exploring beyond knowledge measurement of the course by both authentic assessment and higher-order thinking (HOTS) level of assessment. Authentic assessment is an evaluation strategy that challenges students to perform the tasks they would be expected to do in real-life situations. Furthermore, Muhaimi et al. (2023) define authentic assessment as testing pedagogical problems for students needed and tends more to the higher-order thinking skills. But that doesn't have to mean that HOTS is always higher priority than lower order thinking (LOTS), as lower-order thinking skills are still critical in teaching students to understand integral concepts. Without LOTS, it may be challenging for students to transition into the higher orders of thinking and problem solving. In other words, the challenge of finding a balance between LOTS and HOTS needs to be met in order that language is learnt holistically. In this aspect, teachers need to understand the cognitive levels of their test items so that fairness and concurrence with curriculum is addressed. The Indonesian curriculum also highlights the infusion of HOTS, so teachers must make items in those tests to be more difficult. It is in this context that examining the equilibrium between LOTS and HOTS when developing teacher-made tests becomes essential.

Besides, some teachers exploit materials this not taking into account the kind of tasks (which in turn are used) and if those do or do not promote meaningful communication. This scenario leads to the fact that teachers tend to depend heavily on textbooks without verifying if the tasks are opportunities for language development. Despite the government's efforts to train several programs, research shows Indonesian teachers as a speaker and writer of English are considered moderate. Furthermore, Rozimela et al. (2024) found that the scores of teachers' productive skills are approximately between 68 and 72. This implies that there is a need for ongoing development. Thus, professional development has to include continued education, working together and mentorship for training of teachers. Furthermore, there is potential for mutual assistance between senior and younger/less experienced teachers to help share knowledge and narrow the gap in assessment literacy. The ongoing enhancement of classroom teaching

and learning is crucial because assessment procedures significantly influence the instructional needs of students. In addition, assessment literacy enables teachers to make sound choices about pre-existing test items or may lead to the creation of their own. Therefore, improving the quality of teaching and learning is emphasized to promote teachers' assessment ability.

Other researchers have also confirmed technical deficiencies of teacher-constructed tests that may compromise their validity and utility. For instance, Sukendra (2023) found that only 55% test items are in line with the syllabus and the rest did not represent the signification learning purpose. This observation indicates that even many items are not 'low stakes' in the sense of being critical competences against which anyone would want to be measured. Additionally, a number of test items include errors in grammar or confusing distractors that are likely to cause confusion among learners and impact their performance. Similarly, Gashaye et al. (2022) identified, low content validity in teacher-made tests for the fact that some language skills were more than others not equally tested. Other investigations, such as Apriana et al. (2022) and Utami et al. (2019), found that teacher- made tests generally concentrate on LOTS only resulting very few HOTS questions. This is against the principles of 2013 curriculum, which promotes critical thinking. Furthermore, Rohmah (2019) discovered low discrimination index and inappropriate alternatives in lots of test items. Thus, test weakness still stands as an important factor that should be taken into account in language teaching.

In this regard Sujana et al. (2019) contend that quality of English Education in senior high schools can be enhanced when teachers are empowered with content, process, assessment and professional development. While existing studies on the topic have tended to focus on validation as the completion of test items when they are developed, the present study takes a novel direction by investigating how teachers actually think and do with respect to validity during item-writing. This change is significant because the examination of process can foster a richer understanding about teachers‟ assessment literacy. In addition, knowledge of teachers' conceptions of validity may provide insights into the difficulties that teachers experience when they develop assessments. Furthermore, this study serves to determine how teachers incorporate curriculum guides and references during the construction of their tests. The probe has the potential to be used in training programs by identifying areas where teachers show a lack of knowledge. Furthermore, this study sheds light on the actual conditions taking place in English MGMP communities in West Lombok. Hence, the results can contribute to informing policy and teacher education programmes. Therefore, the findings of this research add up to the knowledge regarding validity practice in teacher-made tests in Indonesian context.

METHODS

This is a qualitative descriptive study, which seeks to provide rich description of how certain phenomena manifest in the eyes and experiences of persons or communities within their own local context. This methodology is beneficial because it enables researchers to collect experiences as they occur naturally, not interfering or altering them. Quoting Lambert and Lambert (2012), there is a way', in which qualitative descriptive research reports results without their being immersed in analytical summaries, so that readers can "see" what happened quite transparently (p. Creswell (2009) also posits that qualitative research aims to investigate and comprehend the meanings people have ascribed to a social problem or issues. Because of this emphasis, the approach allows researchers to explore real-world perceptions, activities and behaviors that could be missing from quantitative designs. This study uses case study to examine how the MGMP English teachers on West Lombok make the test item. A case is warranted as it enables the researcher to observe the goings-on of how a thing works, in its happening. Case studies are particularly appropriate when the links between the phenomenon and its context have not been clearly established (Yin, 2013).

So this model begins to address how teachers, in fact, mediate their material conditions when constructing assessment. Through this integration of qualitative descriptive and case study, thereby the research aims to offer a nuanced portrayal of test-making activities among English MGMP teachers in West Lombok. The research was conducted in several high schools in West Lombok in September 2025.

The sampling approach in this study was a snowball recruitment procedure which is if the first participant meets inclusion criteria they can recommend to another one. Johnson (2014)) and Naderifar et al. (2017) note snowball sampling is useful for guiding the researchers to individuals who possess certain, relevant traits. The focus of this research was on the English teachers for MGMP in high school level in West Lombok. These teachers were selected because they have the habit of preparing test items for their students, to enable them serve as informants on how to develop tests. By the snowball approach, they referred each other, and increased the sample size slowly. A total of ten teachers were ultimately chosen as participants of the study. Their participation was vital, as they provided firsthand experiences with test development procedures and problems. Snowball sampling also promoted the fact that participants were indeed informants about the research issues. Therefore, the sample well represented the requirements of the characteristics for examining test-making practices in the MGMP community.

The data were anlyzed using triangulation through documents and interviews which provided in-depth and credible responses. I opted for interviews as this method enables informants to explain in some detail their experiences, beliefs and practices as described by Hox et al. (2005). Interview questions in this study were formulated from the validity theory to meet the necessary aspects of test quality. At the same time, the semi-structured nature allowed for new ideas to arise freely. Semi-structured interviews provided a combination of structure and flexibility, allowing the researcher to follow up on meaningful points as participants reported them (Kvale & Brinkmann, 2015). Each interview lasted approximately 30 min, was audio-recorded with consent, and centred on the participants' understandings, perspectives and experiences of producing test items. Analyses of documents supplemented the interviews and included examples relevant to teacher-made test items. Creswell & Creswell, 2023). Documentation such as public and/or private records is important sources to corroborate and enhance interviews findings. Thus providing test documents helped validate or clarify participant-reported practices. By taking this joint approach, the investigation was able to capture a global view of what teachers believe about validity and how they apply it in test construction.

For the purpose of ensuring overall quality and trustworthiness of the study, various strategies consistent with Lincoln and Guba (1985) were used. Reliability was confirmed via member checking, where the participants rechecked the information presented to them. Engagement with the research setting over time also increased credibility, as it allowed the researcher to become increasingly familiar with the local activities. Triangulation was another factor in enhancing the credibility and validity of results, as findings were confirmed through checking with several data sources. Transferability was facilitated by rich descriptions of the research setting and participants, so that readers could determine whether or not the findings might be transferable to other settings. Reliability was assured through an audit trail which tracked decisions made during the conduct of the study. Reflexive journaling allowed the researcher to consider potential bias thereby increasing confirmability. Another level of objectivity was introduced in the form of peer debriefing sessions that included input from peers. The careful combination of these strategies ensured the high methodological quality of study. These initiatives helped maintain reliable and evidence-based findings that were based on systematically collected data.

Data analysis procedures were guided by the thematic analysis approach by Miles et al.

(2014), which emphasizes that qualitative data analysis is cyclical and begins during the initial phase of data collection. Data reduction was the first step in this study, entailing summarisation and pasting of extracts into a manageable code. Irrelevant text was then excluded to emphasize the main themes. The second phase was the re-arranging of condensed data in visual or written forms such as matrices and charts, thematic group formation. These views facilitated the recognition of connections and patterns between cases. Once the displays were constructed, the researcher progressed to drawing conclusions and verification. At this point patterns were discerned, explanations fashioned, and interpretations checked against the raw data for believability. Alternative views were also considered, in order not to bias the interpretations too strongly. Verification was an important way to make sure that whatever conclusions were generated were based on evidence, not just guesses. In this iterative and reflective process of analysis, reporting was both data driven and analytically firm.

FINDINGS AND DISCUSSION

Findings

In the process of developing English language tests, teachers demonstrated an awareness of the importance of test validity. Most teachers expressed that their test design was guided by various official documents, such as the national curriculum, lesson plans, syllabi, and instructional modules. These references served as the foundation to ensure that test items aligned with learning objectives. Teachers acknowledged that a well-aligned test helps measure students' competencies effectively. They believed that instructional goals should be reflected in assessment tools to maintain both relevance and fairness. Many teachers emphasized that student ability levels also had to be considered when formulating test items. This consideration ensured that assessments were not too difficult or too easy, allowing for equitable evaluation. Interestingly, teachers did not work in isolation when designing tests. They regularly consulted with colleagues to share ideas and align interpretations of curriculum standards. This peer collaboration was seen as essential to maintaining consistency and improving test quality. The teachers also mentioned the use of test blueprints or grids to map out item indicators and skills being measured. These steps were part of a systematic effort to uphold content validity in assessment practices.

When asked about strategies they used, teachers reported several methods to ensure that their test items were valid and appropriate. They began by referring to documents such as the ATP (Alur dan Tujuan Pembelajaran), curriculum, and syllabi to guide their decisions. These resources helped them create assessments that aligned with learning goals. Teachers also reviewed modules and instructional materials to ensure that the content of the test reflected what had been taught. In addition to this, some teachers conducted try-outs of their tests. The purpose of these trials was to gather feedback and adjust the level of difficulty accordingly. Technological tools also played a role in their strategy. For example, some teachers used the PIJAR application to build and manage question banks. This helped them keep track of test items and maintain consistency across semesters. Furthermore, WhatsApp groups were utilized for peer discussions and collaborative review. These virtual communities allowed teachers to refine their test items with input from others. Despite these efforts, teachers did not mention using more advanced statistical measures, such as discrimination index or reliability testing. This suggests that while they followed some systematic procedures, there was room for growth in using psychometric tools to evaluate item quality.

Despite their efforts, teachers faced a variety of challenges in test development. One of the most common issues mentioned was the varying levels of student ability. Teachers noted that students in the same grade often had differing skill levels, which made it difficult to design tests that were fair for all. Some reported struggling to create items that were accessible to weaker students while still challenging for stronger ones. Another problem was related to ambiguous or unclear guidelines in national documents. For example, instructional

indicators were sometimes interpreted differently by different teachers. In addition, some tests did not align properly with these indicators, reducing their content validity. Teachers also noted that some test items were too difficult or used vocabulary unfamiliar to students. The use of both Bahasa Indonesia and English in a single item also created confusion. Moreover, time constraints posed a significant issue. Teachers were often overwhelmed with responsibilities outside of test creation, leaving little time for thoughtful item development. They also admitted having difficulty developing HOTS (Higher Order Thinking Skills) questions. Lastly, the lack of regular MGMP (English Teacher Working Group) meetings in West Lombok reduced opportunities for professional collaboration and peer evaluation. This absence of external feedback made it harder to improve and validate their tests consistently.

To support interview findings, document analysis was conducted on final exam materials from two schools: SMAN 1 Narmada and SMAN 1 Gerung. At SMAN 1 Narmada, a total of 30 test items were analyzed, including 25 multiple-choice and 5 essay questions. The content validity analysis revealed that 17 items (56.67%) were considered highly appropriate and well-aligned with the ATP. Five items (16.67%) were deemed appropriate, while 8 items (26.67%) did not align with the learning objectives. This suggests that while the majority of items were valid, there were still a notable number of questions that lacked alignment with instructional goals. The cognitive validity analysis further showed that 57% of the items targeted lower-order thinking skills, specifically C1 (Remembering) and C2 (Understanding). About 27% of the items aimed at C3 (Applying) and C4 (Analyzing), representing moderate cognitive engagement. However, higher-order skills like C5 (Evaluating) and C6 (Creating) were underrepresented, with only 10% targeting these domains. This data indicates a tendency to emphasize factual recall over analytical or creative thinking, which limits the test's ability to measure a wider range of student competencies.

In contrast, the final exam at SMAN 1 Gerung consisted of 45 multiple-choice questions. Analysis of these items showed that 42 (93.33%) were highly appropriate and aligned with the curriculum. Only 3 items (6.67%) were considered misaligned. This reflects a strong adherence to content standards and a greater level of consistency in test construction. The cognitive level distribution was more balanced compared to SMAN 1 Narmada. About 31% of items were at the C4 (Analyzing) level, which implies a strong focus on students' ability to examine and break down texts. Another 27% of items were classified under C2 (Understanding) and C3 (Applying), indicating an emphasis on comprehension and contextual use. Surprisingly, only 2% of items focused on C1 (Remembering), showing limited use of rote memory-based questions. The C5 (Evaluating) level appeared in 4% of the items, while none were found in the C6 (Creating) category. This absence of productive tasks implies that while students were asked to interpret and analyze, they were not given opportunities to design, present, or generate new ideas. Overall, SMAN 1 Gerung showed stronger content and cognitive validity than SMAN 1 Narmada, though both schools showed potential areas for growth, particularly in incorporating higher-order thinking tasks

Discussion

The study that English teachers in high school have knowledge about the importance of test validity, especially content validity, when constructing test items. The majority of participants believed that test items must be consistent with the curriculum, syllabus or lesson plan in order to be relevant to objective. This and in tone with Brown (2004) that a good test should reflect the instructional goals, evaluate the competencies that are really intended for evaluation. Providing blueprints and test grids - (Notes 8.3) Teachers' use of blueprints and test grids also provides evidence for Miller, Linn and Gronlund's (2009) theory that a set up in the form of test specification tables is crucial in aligning test content with objectives. Peer work by dialogue and consultation was widely reported, which indicate that the process of test construction is a

collective activity rather than individual behavior. This joint endeavor is consistent with Stiggins (1994) who argued for the necessity of having various perspectives on testing in test f development to increase reliability and decrease bias. Use of testing blueprints also aligns with Nitko and Brookhart (2011) recommendation for clear representation of the relationship among test indicators to enhance the quality of an assessment. Nevertheless, the extensive concentration on recall and the comprehension levels (C1 and C2) at Bloom's taxonomy indicate that teachers remain biased towards lower order thinking skills. This pattern is also supported by the findings of Brookhart (2010), that, in the absence of effective training, many teachers fail to craft assessments that effectively promote critical and creative thought on behalf of their students. So, while they employ some procedures for validity, teachers' uses of assessment need to move closer to higher-order thinking skills.

The test design approach described by teachers reflects a systematic consideration rooted in official surveillance documents and materials. Instructors used ATP, syllabi, modules and lesson plans to align student learning artifacts (McMillan, 2011), as McMillan streamlined sound assessment which starts with clear instructional alignment. Try-outs and adjustment The use of (1998b: 38) complexity clearly reflects a few teachers testing their items in some informal way for appropriateness or difficulty15 (Popham, Developmental evidence as to whether the participants understood what was required when they read items. The use of the PIJAR app and several WhatsApp groups also suggests a tendency toward the integration of technology and peer collaboration in test construction. This confirms Chappuis (2015), who calls the importance of digital tools for enhancing assessment practice. Despite this, absence of reference to psychometric analyses (relability coefficients or item discrimination indexes) points out a lack in teachers technical assessment competences. "It is important for items to work in the same way across students" (Nitko & Brookhart, 2011, p. Without these mechanisms, it is still challenging to assess item quality beyond surface alignment. However, cooperation can ensure the integrity of content, and then more in-depth analysis is required to make a well-grounded evaluation. Teachers' use of informal peer feedback, while beneficial, are unlikely to supersede more robust validation supported by assessment theory. As a result of this reality, their current practices resemble more of a base level validity practice and consist of no further analysis capability in which to tune test item quality accordingly.

In spite of these efforts, teachers continue to encounter significant obstacles that restrict the development of legitimate assessment practices. One big challenge is students' unequal language proficiency, which makes creating fair test items difficult. This concern is grounded in Messick's (1989) conceptualisation of construct validity, which cautions that test devices should not be biased towards or against particular groups of students. When unknown vocabulary and bilingual (Indonesian-English) test items are introduced, it adds to the confusion and possibly disguises students' real level of competence. Moreover, discrepancies in the interpretation of instructional indicators reveal a more systemic problem in the curriculum guidance. The absence of routine MGMP meetings and the lack of active evaluators also undermined the supporting system for test validation. This directly opposes Shepard's (2000) claim that assessment systems need supporting professional communities and feedback mechanisms. Teachers also reported challenges in framing HOTS questions, and needed more training to devise evaluative and creative test items. This question relates to Anderson and Krathwohl's (2001) taxonomy, which defines evaluating (C5) and creating (C6) as critical elements of the spectrum, but also of contemporary assessment. In addition, lack of time and workload were reported, thus supporting the previous accounts of Stiggins (2005) on how capacity-building among teachers needs to be institutionally facilitated for assessment practices to be enhanced. Indeed, these barriers show that, while teachers strive to design valid assessments, systematic and professional obstacles impede the ability to

routinely apply best practices.

CONCLUSION

The findings revealed that most English teachers ensured the content validity of their test items by referring to official documents such as the curriculum, teaching modules, and test blueprints during the test development process. These references helped maintain alignment between the assessments and learning objectives. However, several items still deviated from the materials taught, and none of the teachers addressed whether the test formats appropriately matched the language skills being assessed—such as speaking, writing, listening, or reading. In terms of strategies, teachers adapted test items to students' proficiency levels and collaborated with peers. They also used digital tools like WhatsApp and the PIJAR app to manage test banks efficiently. Despite these efforts, none of the teachers reported using analytical tools such as the discrimination index or discrimination power to evaluate their test items. The main challenge teachers encountered was the wide range of students' abilities, which made it difficult to design fair assessments suitable for all learners. Many students were also performing below their expected academic level. This situation was further compounded by the lack of recent professional development activities or coordination from the English MGMP in West Lombok in the past one to two years.

ACKNOWLEDGMENT

I sincerely thank to Allah S.W.T for the strength and opportunity to complete this research. I also thank my supervisors for all of their guidance and support, and the examiner for the valuable critique. Furthermore, an acknowledgment is due to the respondents for their participation. Finally, thank you to all others who contributed to this research.

REFERENCES

Anderson, L. W., & Krathwohl, D. R. (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition. Addison Wesley Longman, Inc.

Apriana, S., Lubis, A. A., & Sofyan, D. (2022). HIGHER-ORDER THINKING SKILLS ON TEACHERS-MADE TESTS BY ENGLISH TEACHERS OF A SENIOR HIGH SCHOOL IN BENGKULU CITY. In FKIP Universitas Lambung Mangkurat Banjarmasin (Vol. 5).

Brown, D. H., & Abeywickrama, P. (2019). LANGUAGE ASSESMENT Principles and Classroom Practices.

Creswell, J. W. (2009). John W. Creswell-Research Design_ Qualitative, Quantitative, and Mixed Methods Approaches-SAGE Publications, Inc (2009).

Creswell, J. W., & Creswell, J. D. (2023). Research Design: Qualitative, Quantitative, and Mixed Methods Approaches Sixth Edition.

Gashaye, S., Degwale, Y., & Ma, (. (2022). The Content Validity of High School English Language Teacher Made Tests: The Case of Debre Work Preparatory School, East Gojjam, Ethiopia. http://indusedu.org

Johnson, T. (2014). Snowball Sampling: Introduction.https://doi.org/10.1002/9781118445112.STAT05720

Kumar, R., & Krishan Lal, M. (2024). TEACHER MADE TEST (Vol. 12). www.ijcrt.org

Lambert, V. A., & Lambert, C. E. (2012). Editors: Pacific Rim International Journal of Nursing Research. In Pacific Rim Int J Nurs Res.

Lissitz, & Robert W. (2009). The Concept of Validity Revisions, New Directions, and Applications. www.infoagepub.com

Muhaimi, L. ., Sahuddin, S., Fitriana, E. ., & Azis, A. D. . (2023). The Implementation of Authentic Assessment in the Literary Subjects: A Pedagogical Perspective. Jurnal Ilmiah Profesi Pendidikan, 8(1), 193–200. https://doi.org/10.29303/jipp.v8i1.1173

Naderifar, M., Goli, H., & Ghaljaie, F. (2017). Snowball Sampling: A Purposeful

Method of Sampling in Qualitative Research. 14. https://doi.org/10.5812/SDME.67670

Ngafif, A., Sukarni, S., Nugraeni, I. I., Sahidah, N., Pithaloka, K. D., & Ningsih, S. C. (2022). Factors affecting the reliability of senior high schools' english teacher-made test in Kebumen. Jurnal Pendidikan Surya Edukasi (JPSE), 8(2), 162–177. https://doi.org/10.37729/jpse.v8i2.2164

Risma, R. (2023). A THESIS AN ANALYSIS OF THE ENGLISH TEACHER-MADE TEST FOR SECOND GRADE STUDENTS AT UPT SMKN 3 PAREPARE.

Rohmah, N. (2019). Validity and Reliability Study on Teacher-Made Assessment for English Mid-Term Examination.

Rozimela, Y., Fatimah, S., Adnan, A., & Tresnadewi, S. (2024). EFL teachers' perceived productive skills for effective teacher professional development program. Studies in English Language and Education. https://doi.org/10.24815/siele.v11i3.37277

Satria Elmiana, D. (2018). The Journal of Asia TEFL A Critical Analysis of Tasks in Senior High School EFL Textbooks in Indonesia. THE JOURNAL OF ASIA TEFL, 15(2), 462–470. https://doi.org/10.18823/asiatefl.2018.15.2.14.462

Sujana, I. M. (2000). Movements in Language Testing: From Grammar-Based to Communicative Language Testing. Jurnal Ilmu Pendidikan FKIP UNRAM, 13(48), 1-8.

Sujana, I. M. (2016). Assessing Oral Proficiency ASSESSING ORAL PROFICIENCY: Problems and Suggestions for Elicitation Techniques.

Sujana, I. M., Andayani, Y., Baidowi, B., Ilhamdi, M. L., Suryanti, N. M. N., & Mariâ€™i, M. (2019). ANALISIS HASIL UJIAN NASIONAL BAHASA INGGRIS SMA DAN PENGEMBANGAN MODEL PENINGKATAN MUTU PEMBELAJARAN BAHASA INGGRIS DI PROPINSI NUSA TENGGARA BARAT. Jurnal Ilmiah Profesi Pendidikan, 1(2). https://doi.org/10.29303/jipp.v1i2.9

Sukendra, I. (2023). ITEM ANALYSIS OF ENGLISH TEACHER-MADE TEST.

Umar, I., Akib, E., & Asrianto Setiadi, M. (2022). An Analysis of The Validity of English Test Made by The Teacher in SMP Negeri 1. Bontomarannu Journal of Language Testing and Assessment, 2(1), 47.

Utami, F. D., Nurkamto, J., & Marmanto, S. (2019). Higher-Order Thinking Skills on Test Items Designed by English Teachers: A Content Analysis AR TI CL E IN FO AB STR A CT. www.ijere.com

Yin, R. K. (2013). Case Study Research Design and Methods (Applied Social Research Methods) (Robert K. Yin) (Z-Library).