Jurnal Review Pendidikan dan Pengajaran
http://journal.universitaspahlawan.ac.id/index.php/jrpp
Volume 7 Nomor 4, 2024
P-2655-710X e-ISSN 2655-6022

Submitted : 29/08/2024
Reviewed : 02/09/2024
Accepted : 06/09/2024
Published : 10/09/2024

Siti Danti[1]
M. Adib Nazri[2]
Zahratul Fikni[3]
Laila Wati[4]

AN ANALYSIS OF TEACHERS' SUMMATIVE ASSESSMENT: EXAMINING THE QUALITY OF TEACHER-MADE TEST

Abstrak

Penelitian ini bertujuan untuk menganalisis kualitas tes yang dibuat oleh guru dalam penilaian sumatif siswa. Pendekatan kuantitatif deskriptif digunakan dalam penelitian ini dengan 30 peserta dari MA Hamzanwadi Pancor di kelas dua belas. Teknik sampling purposive digunakan untuk memilih peserta yang sedang mengikuti ujian akhir mereka. Untuk mengumpulkan data, peneliti mengambil lembar ujian, lembar jawaban, dan kunci jawaban siswa dalam bentuk soft copy dan digital, kemudian dianalisis menggunakan MS Excel 2016. Hasilnya menunjukkan empat aspek untuk menentukan kualitas tes yang dibuat oleh guru, yaitu indeks kesulitan, indeks diskriminasi, validitas, dan uji reliabilitas. Hasil indeks kesulitan menunjukkan bahwa 33,66% soal dapat diterima, 15,30% terlalu sulit, dan 2,4% terlalu mudah. Indeks diskriminasi mengungkapkan bahwa 14,28% soal memiliki diskriminasi yang baik, sementara 36,72% memiliki diskriminasi yang buruk. Hasil validitas menunjukkan bahwa 19,38% soal valid, dan 31,62% tidak valid. Reliabilitas tes adalah 0,7, yang diklasifikasikan sebagai "tinggi." Kesimpulannya, meskipun tes menunjukkan indeks kesulitan dan reliabilitas yang baik, indeks diskriminasi dan validitas memerlukan perbaikan untuk penilaian kemampuan siswa yang lebih akurat.

Kata kunci: Penilaian Summative, Tes Buatan Guru, Kualitas Tes

Abstract

This study aimed to analyze the quality of teacher-made tests in students' summative assessment. Descriptive quantitative was used in this study with 30 participants from MA Hamzanwadi Pancor at twelveth grade. Purposive sampling was used to choose the participants that currently engaged their final exam. In order to collect the data, the researcher took paper test, students' answer sheets, and the key answer in soft copy and digital form then it was analyzed using Ms. Exel 2016. The result showed that there were four aspects to determine the quality of teacher-made test such as difficulty index, index of discrimination, validity and reliability testing. The difficulty index results showed 33;66% applicable questions, 15;30% too difficult, and 2;4% too easy. The discrimination index revealed 14;28% of questions had good discrimination, while 36;72% had poor discrimination. Validity results indicated 19;38% of questions were valid, and 31;62% were invalid. The test's reliability was 0.7, classified as "high." In conclusion, the test had a good difficulty index and reliability, but the discrimination index and validity need improvement for a more accurate assessment of student abilities.

Keywords: Summative Assessment, Teacher Made-Test, Test Quality

INTRODUCTION

Assessment is the process of gathering, analyzing, and evaluating information about the abilities, knowledge, skills, or performance of individuals, groups, or other entities (d'Amato & Hunter, 2024) . The primary purpose of assessment is to measure the extent to which someone or something has achieved set goals or reflects the capabilities they possess. Assessment can be

[1,2,3,4)] Program Studi Pendidikan Bahasa Inggris Fakultas Bahasa, Seni, dan Humaniora Universitas Hamzanwadi
email: sitidanti1215@gmail.com[1], adibnazri88@gmail.com[2], zahratulfiknii@gmail.com[3], ladyazzurry@gmail.com[4]

conducted in various contexts, including education, employment, health, or various other fields. In education, a kind of assessment is called summative and formative assessments (Chong & McArthur, 2023).

According to Yan & Brown, (2021) summative assessment is different from formative assessment that is focus on the evaluation process conducted during the learning. Meanwhile, summative assessment is an evaluation method or test that is typically used to measure a student's overall understanding and knowledge of a subject at the end of an instructional period, such as a course, unit, or school year (Utami et al., 2020) . It is used to determine what a student has learned and achieved as a result of their learning experiences and to assign a final grade or. It takes various forms, including exams, final projects, standardized tests, or other culminating assessments that assess the extent to which students have met the learning objectives and standards set for the educational program (Thambusamy & Singh, 2021).

Teachers hold responsibility for designing assessments that accurately assess student learning outcomes and provide meaningful feedback (Driscoll & Wood, 2023). However, the process of creating high-quality summative assessments is complex and multifaceted. It involves careful consideration of learning objectives, content coverage, types of questions, quality of items, and the overall structure of the test (Ahmad, 2020). Understanding the factors that influence the quality of teacher-made assessments is critical to improving overall assessment practice in educational settings.

The quality of teacher-made tests serves as the linchpin of equitable and precise student assessment (Yusup, 2021). A meticulously designed test should not only harmonize seamlessly with the prescribed syllabus and core competencies but also be meticulously tailored to accommodate diverse learning styles, fostering a comprehensive understanding of the material (Stevens & Levi, 2023). In this digital era, teachers often take questions from the internet without checking the alignment with the material. It causes the confusing students and leave them unprepared for exams, ultimately making the learning process ineffective. Furthermore, it should consistently yield dependable results while maintaining high validity to serve as a dependable yardstick for evaluating students' educational progress (Winstone & Boud, 2022).

In dealing with this problem, it is important for teachers and instructors in Indonesia to choose summative questions carefully. Validation of questions and the use of variations in evaluation methods need to be a focus in the student assessment process (Jackson, 2019). In addition, the teacher's role in developing summative questions that are in accordance with students' abilities is also very important. Thus, the use of summative questions can be more effective and provide more reliable test results to improve the quality of education in Indonesia. The novelty of this research is the educational observer can gain the informations from this study that the summative assessment need more pay attention especially in many school of Indonesia because the spread of the quality in Indonesia is different.

By addressing these issues, this research aims to analyse the quality of teacher-made tests in students' summative assessment at the twelveth-grade in MA Hamzanwadi Lombok. The researcher will do the test using teacher made-test on student's final exam with 50 multiple-choice. A thorough analysis of teacher-made tests can provide valuable insights for educational policymakers, schools, and teachers to refine assessment practices. It will foster a more equitable and effective learning environment.

METHOD

This research design employed a descriptive quantitative approach. Descriptive quantitative research involved the systematic collection and analysis of numerical data to objectively depict and summarize the characteristics of a population, group, or phenomenon under study, particularly in the context of teacher-made tests (Indriyani, 2024). The collected data was often used to create visual representations and draw generalizations about the broader population, emphasizing the importance of ensuring validity and reliability in the research process.

The setting of the study referred to the place and time to conduct the study. This study was conducted in 2024 at MAH Pancor in Pancor, East Lombok, during the final test of the teaching and learning. The target population of this study was the twelveth grade students of

MA Hamzanwadi in the school year 2023-2024. The total number of the population was 30 students. A sample consists of a selected group of elements chosen from a defined population. The study focused on the twelveth-grade class and their teacher as the target to analyse the quality of teacher-made tests.

In this research, the instrument that researcher use is the students answer sheet which student gave their respond to English summative in the soft copy data form. Moreover, the researcher take the answer key that writer used was the answer key based on teacher-made. Then, the English summative test paper that the writer used was the odd semester of 12th grade of MA Hamzanwadi in academic year 2024. The total number of the item test were 50 questions, and all of the questions were multiple-choice item.

In order to collect the data, the writer visited the school to request the English summative paper test, students' answer sheets, and the key answer in soft copy and digital form. This involved directly communicating with the English teacher at the school to request copies of the exam, which included the test papers used by the students, the answer sheets filled out by the students, and the key answer in digital format. This way, the writer can obtain the exam materials and the necessary data for further analysis. Furthermore, the document analysis in this study used the test items, students' answer sheet, and key answer from the teacher. The test items checked to analyse the validity and reliability of the teacher-made test. It was the proof whether the items were appropriate to what the students had learned in their classroom. Additionally, Exel was chosen to analyse the data in this research.

RESULT AND DISCUSSION

In this part, the researcher explained about the result of the study after calculating the data from the teacher-made test on the students result at their final exam. The researcher used Exel 2016 to count the data. As the result, there were four aspects to determine the quality of teacher-made test. Those were difficulty index, index of discrimination, validity, reliability testing. Here were the detail explanations:

1. Difficulty index

The difficulty index was the percentage of proportion of students who answered the item correctly. The higher the percentage of the students who answered correctly, the easier the item was. Here were the results of the difficulty index based on teachers' made-test on the students' final exams results.

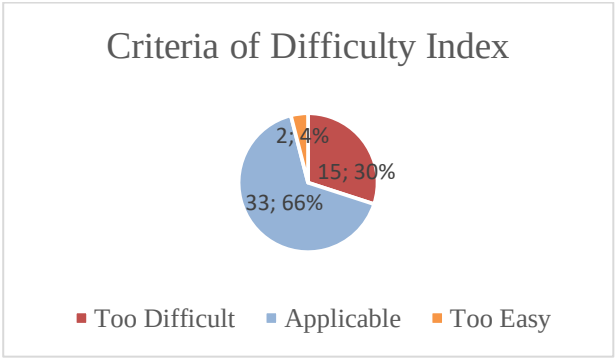

Chart 1.

Based on the data in the chart above, there were three analysis results based on students' answers to the summative test created by the teacher. Out of 50 questions, there were three criteria: too difficult, applicable, and too easy. The number of applicable questions was 33 questions or 66%, while the number of too difficult questions was 15 questions or 30%, and the number of too easy questions was only 2 questions or 4%. Based on this analysis, the difficulty index of the teacher-made test showed that most of the questions were in the applicable category, which meant the test was well-designed as it was neither too difficult nor too easy based on the students' response percentages.

2. Index of Discrimination

The discrimination index was a measure used to evaluate how well the test can distinguish between students with high and low understanding or ability. Its functions included identifying effective questions, improving test quality, measuring question validity, enhancing learning assessment, and optimizing teaching methods. Questions with a high discrimination index indicated effectiveness in assessment, while questions with a low index needed to be revised or removed. Thus, the discrimination index helped ensure that tests were fair, accurate, and effective in measuring students' abilities. Look at the result from the chart.

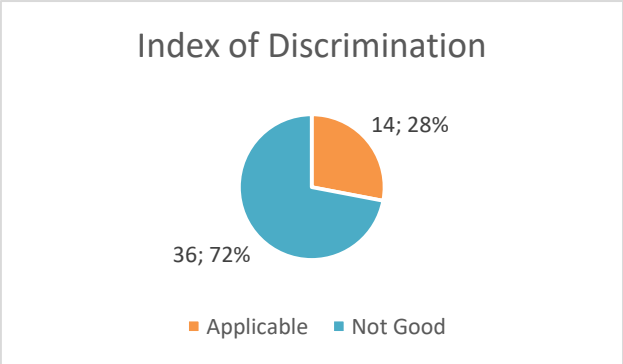

Chart 2

The data from the students' answers showed that out of 50 questions, 14 had a good discrimination index (applicable) and 36 had a poor discrimination index (not good). Only 14 out of 50 questions can effectively distinguish between students with high and low abilities, meaning that only 28% of the questions accurately measured the difference in students' abilities. Conversely, 36 out of 50 questions, or 72%, failed to distinguish well between high and low ability students. These questions might be too easy, too difficult, or unclear, thus not providing useful information about the differences in students' abilities.

3. Validity Testing

Validity was the measure of how well a test or assessment instrument measures what it was supposed to measure. A test was considered valid if its results can be trusted and were relevant to the measurement purpose.

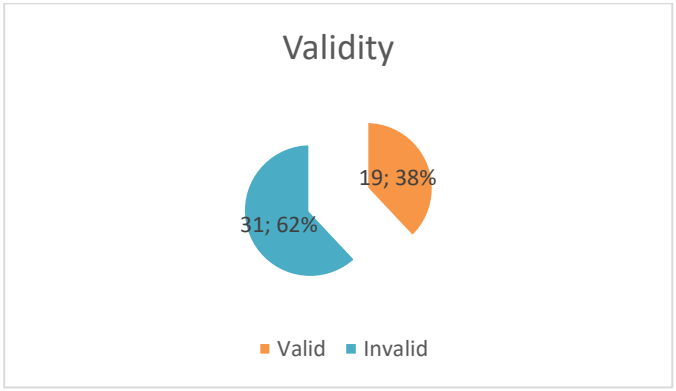

Chart 3.

The validity test results from students' answers showed that out of 50 questions created by the teacher, 19 questions were deemed valid and 31 questions were deemed invalid. Out of 50 questions, only 19 are valid, meaning that only 38% of the questions accurately and relevantly measured what they were supposed to measure. These questions provided reliable information about the students' abilities or knowledge being tested. Conversely, 31 out of 50 questions, or 62%, were found to be invalid, indicating that the majority of the questions fail to accurately measure the learning objectives.

4. Reliability Testing

Reliability was the measure of the consistency and stability of a test's results over time. A reliable test produced similar outcomes under consistent conditions, ensuring that the results were dependable and repeatable. In this research, after calculating the item variance, the researcher obtained a total item variance of 32. The reliability calculated from this result was 0.7, which, according to experts, fell into the "high" category.

In conclusion, the teacher-made test showed that out of 50 questions, 19 were valid and 31 were invalid, indicating that only 38% of the questions accurately measured what they were supposed to. Despite this, the overall reliability of the test was 0.7, which was considered "high" according to experts, suggesting that the test results were consistent and dependable. Therefore, while the test demonstrates good reliability, there was a need to improve the validity of a significant portion of the questions to ensure a more accurate assessment of student abilities.

Discussion

The quality of summative assessments crafted by teachers is critical in accurately measuring student understanding and guiding instructional strategies. It indicates that well-designed tests typically have a majority of questions with moderate difficulty, ensuring a fair assessment of student performance. (Iqbal & Riaz, 2023). For example, (Murphy et al., 2023) suggests that around 60-70% of test questions should fall into this range. Based on the data from the teacher-made summative test, most of the questions are considered appropriate, indicating that the test is generally well-designed and aligned with recommendations from previous research.

The proportion indicated that some questions are overly challenging, potentially affecting the accuracy of the assessment in reflecting students' true abilities (Panadero & Jonsson, 2020). However, some questions are deemed too difficult, with the number exceeding the ideal range. Teachers may need to conduct a deeper analysis of the questions considered too difficult and identify the factors causing this difficulty. It includes revising the questions to ensure they are fair and representative of the material taught, or perhaps providing more practice on the topics that were most challenging for students before the test is given.

Additionally, the discrimination index is a crucial tool for evaluating the effectiveness of test questions in distinguishing between students with varying levels of understanding or ability (Olipas & Luciano, n.d.). It serves multiple functions, such as identifying effective questions, improving overall test quality, measuring the validity of questions, enhancing learning assessments, and optimizing teaching methods. High discrimination index questions are indicative of effective assessment, accurately reflecting the abilities of students.

The teacher-made test reveals significant issues with the discrimination index of the questions. A small portion of the questions effectively distinguishes between students with high and low abilities, while the majority fails to do so (Urhahne & Wijnia, 2021). These poorly performing questions may be too easy, too difficult, or unclear, thus not providing useful information about the differences in students' abilities. The high proportion of ineffective questions suggests a need for significant revision to improve the test's ability to accurately assess student performance and enhance its overall quality. Teachers should consider using item analysis techniques to identify and revise or replace questions that do not perform well.

Validity is a fundamental aspect of assessment quality, measuring how well a test or instrument evaluates what it is intended to measure. A valid test produces results that are trustworthy and relevant to the assessment’s purpose. According to (Earle, 2020), the importance of validity in educational assessments indicates that tests with high validity are crucial for accurate measurement and meaningful interpretation of student performance. The high number of invalid questions in this teacher-made test highlights a critical area for improvement.The potential benefit of professional development in test design and validation processes for educators (Bragg et al., 2021; Sancar et al., 2021; Yurtseven Avci et al., 2020). Providing teachers with training on creating valid assessment items and employing validity analysis tools can enhance the overall quality of assessments.

Reliability measures the consistency and stability of a test's results over time, ensuring that outcomes remaine dependable and repeatable under consistent conditions (Ryan et al., 2021). In the conducted research, the reliability of a teacher-made test has been assessed

through the calculation of item variance, resulting in a total item variance of 32. This calculation yields a reliability coefficient of 0.7, categorizing it as "high" according to expert standards. High reliability indicates that the test consistently generates similar outcomes across various administrations, underscoring its reliability as a robust tool for assessing student performance. Such reliability holds significant importance for educators, as it guarantees that assessment results are stable and can be confidently used to gauge students' true abilities over time (Zuhriyah,2020).

Understanding the reliability of assessments is critical for educators as it directly impacts the trustworthiness and utility of the data gathered from tests. A high reliability coefficient, such as the one found in teacher-made tests, suggests that the results are likely to be stable and consistent over repeated administrations (Murphy et al., 2023). This stability enables educators to confidently use assessment data to monitor student progress, identify learning gaps, and tailor instructional strategies accordingly. Moreover, emphasizing methods like item variance calculations to assess reliability underscores the importance of employing rigorous evaluation techniques in educational assessment practices. These approaches not only validate the integrity of assessment results but also contribute to enhancing overall educational effectiveness by ensuring that assessments reliably measure what they intend to assess.

Reliability assessment highlights the teacher-made test's capability to produce dependable and consistent results, crucial for evaluating student learning effectively (Martin, 2022). By maintaining high reliability, educators can trust that the test accurately reflects students' understanding and progress across different testing instances. This aspect not only enhances the credibility of assessment practices but also supports informed decision-making in educational settings.

CONCLUSION

In education, teacher-made tests are vital for gauging students' learning progress and guiding instructional decisions. A well-designed test encompasses questions of varying difficulty levels, ensuring fairness for all students. However, in a recent teacher-made test, approximately 30% of the questions were deemed too challenging, indicating areas for improvement. Teachers can enhance their tests by honing their skills in crafting fair questions and utilizing assessment tools to evaluate question quality, thus gaining deeper insights into student learning. Furthermore, it is crucial for tests to effectively differentiate between students' levels of understanding. Yet, in the examined teacher-made test, only a few questions achieved this effectively, suggesting a need for additional training for teachers.

By acquiring new methods and leveraging supportive resources, educators can develop assessments that accurately reflect students' knowledge levels.Tests also must accurately measure the intended learning objectives. However, in the discussed test, only 38% of questions achieved this standard, signaling potential issues in test construction. Teachers can learn from past research and incorporate best practices to enhance the quality of their assessments. Ensuring the reliability of tests allows teachers to gain meaningful insights into students' progress and provide targeted support to foster improved learning outcomes

REFERENCE

Afriyuninda, E., & Oktaviani, L. (2021). THE USE OF ENGLISH SONGS TO IMPROVE ENGLISH STUDENTS'LISTENING SKILLS. Journal of English Language Teaching and Learning, 2(2), 80–85.

Ahmad, Z. (2020). Summative Assessment, Test Scores and Text Quality: A Study of Cohesion as an Unspecified Descriptor in the Assessment Scale. European Journal of Educational Research, 9(2), 523–535.

Bragg, L. A., Walsh, C., & Heyeres, M. (2021). Successful design and delivery of online professional development for teachers: A systematic review of the literature. Computers & Education, 166, 104158.

Chomsky, J. (2021). Language and interpretation. Inference, Explanation, and Other Frustrations: Essays in the Philosophy of Science, 14, 99.

Chong, D. Y. K., & McArthur, J. (2023). Assessment for learning in a Confucian-influenced culture: beyond the summative/formative binary. Teaching in Higher Education, 28(6), 1395–1411.

d'Amato, A. L., & Hunter, S. T. (2024). Creativity training needs assessment for homeland security enterprise: a case for creative thinking. Journal of Policing, Intelligence and Counter Terrorism, 19(1), 61–82.

Driscoll, A., & Wood, S. (2023). Developing outcomes-based assessment for learner-centered education: A faculty introduction. Taylor & Francis.

Jackson, D. A. (2019). Interim Assessments as a Predictive Tool and Driver of Formative Assessment Practices to Improve Student Performance on State Assessments. East Carolina University.

Murphy, D. H., Little, J. L., & Bjork, E. L. (2023). The value of using tests in education as tools for learning—not just for assessment. Educational Psychology Review, 35(3), 89.

Olipas, C. N. P., & Luciano, R. G. (n.d.). Analyzing Test Performance of BSIT Students and Question Quality: A Study on Item Difficulty Index and Item Discrimination Index for Test Question Improvement.

Rustamova, S. (2023). PROBLEMS AND SOLUTIONS TO ENGLISH TEACHING. Журнал Иностранных Языков и Лингвистики, 5(5).

Sancar, R., Atal, D., & Deryakulu, D. (2021). A new framework for teachers' professional development. Teaching and Teacher Education, 101, 103305.

Stevens, D. D., & Levi, A. J. (2023). Introduction to rubrics: An assessment tool to save grading time, convey effective feedback, and promote student learning. Routledge.

Thambusamy, R. X., & Singh, P. (2021). Online Assessment: How Effectively Do They Measure Student Learning at the Tertiary Level? The European Journal of Social & Behavioural Sciences.

Urhahne, D., & Wijnia, L. (2021). A review on the accuracy of teacher judgments. Educational Research Review, 32, 100374.

Utami, A. P., Dewi, E. S., & Paramartha, G. Y. (2020). Summative Assessment of Tenth-Grade English Teachers from Hots Perspective. Jurnal Bahasa Lingua Scientia, 12(2), 295–314.

Winstone, N. E., & Boud, D. (2022). The need to disentangle assessment and feedback in higher education. Studies in Higher Education, 47(3), 656–667.

Yan, Z., & Brown, G. T. L. (2021). Assessment for learning in the Hong Kong assessment reform: A case of policy borrowing. Studies in Educational Evaluation, 68, 100985.

Yule, G. (2022). The study of language. Cambridge university press.

Yurtseven Avci, Z., O'Dwyer, L. M., & Lawson, J. (2020). Designing effective professional development for technology integration in schools. Journal of Computer Assisted Learning, 36(2), 160–177.

Yusup, R. (2021). Introducing Dynamic Testing to Teachers in Malaysia: An Experimental Investigation of Its Effects on Teachers' Beliefs and Practices about Assessment. Durham University.

Afriyuninda, E., & Oktaviani, L. (2021). THE USE OF ENGLISH SONGS TO IMPROVE ENGLISH STUDENTS'LISTENING SKILLS. Journal of English Language Teaching and Learning, 2(2), 80–85.

Ahmad, Z. (2020). Summative Assessment, Test Scores and Text Quality: A Study of Cohesion as an Unspecified Descriptor in the Assessment Scale. European Journal of Educational Research, 9(2), 523–535.

Bragg, L. A., Walsh, C., & Heyeres, M. (2021). Successful design and delivery of online professional development for teachers: A systematic review of the literature. Computers & Education, 166, 104158.

Chomsky, J. (2021). Language and interpretation. Inference, Explanation, and Other Frustrations: Essays in the Philosophy of Science, 14, 99.

Chong, D. Y. K., & McArthur, J. (2023). Assessment for learning in a Confucian-influenced culture: beyond the summative/formative binary. Teaching in Higher Education, 28(6), 1395–1411.

d'Amato, A. L., & Hunter, S. T. (2024). Creativity training needs assessment for homeland security enterprise: a case for creative thinking. Journal of Policing, Intelligence and Counter Terrorism, 19(1), 61–82.

Driscoll, A., & Wood, S. (2023). Developing outcomes-based assessment for learner-centered education: A faculty introduction. Taylor & Francis.

Jackson, D. A. (2019). Interim Assessments as a Predictive Tool and Driver of Formative Assessment Practices to Improve Student Performance on State Assessments. East Carolina University.

Murphy, D. H., Little, J. L., & Bjork, E. L. (2023). The value of using tests in education as tools for learning—not just for assessment. Educational Psychology Review, 35(3), 89.

Olipas, C. N. P., & Luciano, R. G. (n.d.). Analyzing Test Performance of BSIT Students and Question Quality: A Study on Item Difficulty Index and Item Discrimination Index for Test Question Improvement.

Rustamova, S. (2023). PROBLEMS AND SOLUTIONS TO ENGLISH TEACHING. Журнал Иностранных Языков и Лингвистики, 5(5).

Sancar, R., Atal, D., & Deryakulu, D. (2021). A new framework for teachers' professional development. Teaching and Teacher Education, 101, 103305.

Stevens, D. D., & Levi, A. J. (2023). Introduction to rubrics: An assessment tool to save grading time, convey effective feedback, and promote student learning. Routledge.

Thambusamy, R. X., & Singh, P. (2021). Online Assessment: How Effectively Do They Measure Student Learning at the Tertiary Level? The European Journal of Social & Behavioural Sciences.

Urhahne, D., & Wijnia, L. (2021). A review on the accuracy of teacher judgments. Educational Research Review, 32, 100374.

Utami, A. P., Dewi, E. S., & Paramartha, G. Y. (2020). Summative Assessment of Tenth-Grade English Teachers from Hots Perspective. Jurnal Bahasa Lingua Scientia, 12(2), 295–314.

Winstone, N. E., & Boud, D. (2022). The need to disentangle assessment and feedback in higher education. Studies in Higher Education, 47(3), 656–667.

Yan, Z., & Brown, G. T. L. (2021). Assessment for learning in the Hong Kong assessment reform: A case of policy borrowing. Studies in Educational Evaluation, 68, 100985.

Yule, G. (2022). The study of language. Cambridge university press.

Yurtseven Avci, Z., O'Dwyer, L. M., & Lawson, J. (2020). Designing effective professional development for technology integration in schools. Journal of Computer Assisted Learning, 36(2), 160–177.

Yusup, R. (2021). Introducing Dynamic Testing to Teachers in Malaysia: An Experimental Investigation of Its Effects on Teachers' Beliefs and Practices about Assessment. Durham University.