FROM TESTING TO IMPROVEMENT: LEVERAGING QUIZIZZ REPORTS FOR EDUCATIONAL ENHANCEMENT

Hery Yanto The[1]
[1]Institut Nalanda, Jakarta Timur, Indonesia

*Corresponding author: heryyantothe@gmail.com

Abstract
Technology-integrated assessment tools have become essential in bridging pedagogy and data-driven educational decision-making, especially in the digital age. This study investigates how Quizizz, a gamified learning platform, can be leveraged to provide actionable data for assessing both course and student performance. While Quizizz is widely used for real-time testing, its rich reporting features generate extensive datasets that can inform instructional improvements. Data were purposively sampled from the mid-term and final tests of Mandarin 1 and Mandarin 2 courses during the 2023/2024 academic year. Using a quantitative approach, the study systematically retrieved, analyzed, and interpreted Quizizz-generated reports through descriptive statistics, Spearman Correlation, and Kruskal-Wallis tests. The results reveal consistent patterns in student responses, question difficulty, and course alignment with learning outcomes. These findings demonstrate that Quizizz data can effectively support improvements in course content, test design, and targeted student support. Educators can enhance teaching strategies and assessment accuracy by integrating educational technology with rigorous data analytics, fostering reflective practice and continuous improvement in language instruction. This study contributes to the discourse on sustainable digital transformation in education by illustrating how learning analytics grounded in social science pedagogy can drive ongoing enhancement in language teaching.
Keywords: assessment, data analytic, evaluation, testing, quizizz

INTRODUCTION

The increasing integration of digital technologies like Quizizz in educational settings has transformed assessment methods by providing educators with extensive datasets that inform instructional decisions and improve student outcomes (Aulia & Warni, 2024; Nurliana & Nugroho, 2021; Umam et al., 2025). Unlike traditional examinations, Quizizz offers immediate feedback and detailed data that allow instructors to assess student performance, question effectiveness, and alignment with educational goals through descriptive and diagnostic analytics (Munawir & Hasbi, 2021; Zhao, 2019). This data-driven approach enables educators to identify learning deficits, tailor interventions, and continuously improve curriculum and teaching methods (Saepul et al., 2023; Umam et al., 2025). This study assesses the effectiveness of Quizizz as a technology-based evaluation instrument to improve language training by meticulously analyzing the data produced by the platform.

Prior studies have established the motivating and engagement advantages of Quizizz across diverse subjects and school levels. Research has demonstrated that Quizizz's interactive and multimodal elements can inspire students to think critically and improve their performance in STEM fields and language acquisition (Daulay et al., 2023; Mesterjon et al., 2024; Yunus & Hua, 2021). Upon the integration of Quizizz, Henukh et al. (2022) conducted research that demonstrated substantial enhancements in students' motivation and academic performance in the field of physics. The platform's easy-to-use interface and instant feedback are also cited in studies on students' views as important elements that assist students in identifying areas for growth and modifying their learning approaches (Munawir & Hasbi, 2021; Yunus & Hua, 2021). Despite these benefits, most of the research has concentrated on short-term impacts like motivation and engagement, paying little attention to using Quizizz-generated data for methodical development of curriculum and enhancement of instruction.

The current study fills this gap by focusing on in-depth data analytics of reports produced by Quizizz within the scheme of Mandarin teaching. This emphasis signifies an innovative contribution by transcending superficial engagement indicators to facilitate a more thorough examination of student learning practices and the quality of assessments. The study uses simple statistics and tests, like Spearman correlation and Kruskal-Wallis, to look at the relationships between mid-term and final exams and to see how student performance varies across different tests. The Spearman correlation examines the consistency of student performance across related exams, whereas the Kruskal-Wallis test determines the presence of significant performance disparities among the tests, thereby guiding instructional modifications and equitable assessment development.

The purpose of this study is to demonstrate how an in-depth exploration of Quizizz-generated data can provide actionable insights for instructional improvement and continuous curriculum enhancement in language education. By harnessing the platform’s detailed reporting features and applying rigorous statistical methods, educators can gain a nuanced understanding of student performance patterns, question effectiveness, and course alignment with learning objectives. These insights enable data-driven decision-making and targeted support for students who may struggle, ultimately fostering more effective teaching strategies and improved learning outcomes. This research contributes to the broader discourse on sustainable digital transformation in education by showcasing how learning analytics grounded in social science pedagogy can drive ongoing improvements in language instruction.

METHOD

This study uses a quantitative approach to analyze student performance and question difficulty across four Mandarin language assessments. It begins with descriptive statistics to provide overall performance trends and question characteristics. To further investigate student outcomes, Spearman correlation and the Kruskal-Wallis test are applied. Spearman correlation examines whether student achievement is consistent between paired assessments within each course, such as mid-term and final exams. The Kruskal-Wallis test evaluates whether significant differences exist in student performance across all four tests, helping assess the impact of assessment design and timing. Together, these methods offer a comprehensive evaluation of individual and group learning patterns, supporting data-driven insights into curriculum effectiveness and student learning.

Descriptive statistics find trends and outliers in performance data, and Spearman correlation checks to see if scores are stable and true across tests that are similar. Kruskal-Wallis's test, which doesn't assume a specific data pattern, is useful for comparing several groups when the data doesn't meet certain requirements, making sure that the conclusions about differences in student results are reliable. These methodologies are broadly recognized in educational research for delivering objective, evidence-based suggestions that improve instructional practice.

Hypothesis testing guides the analysis. For Spearman correlation, the null hypothesis (H_0) states there is no significant relationship between mid-term and final test scores within each course, while the alternative hypothesis (H_1) claims a significant relationship exists, indicating consistent student performance. For the Kruskal-Wallis test, data were restricted to students enrolled in both Mandarin 1 and Mandarin 2 to ensure complete data sets. Here, H_0 posits equal median performance across all four assessments, whereas H_1 suggests at least one assessment’s median differs significantly. This approach allows a robust examination of performance variation across assessments and courses.

Data was collected from mid-term and final exams for Mandarin 1 and Mandarin 2. Students enrolled in only one course were excluded from the Kruskal-Wallis’s analysis to maintain consistency and avoid bias. All statistical analyses were performed using a social science statistics platform to ensure accuracy and reproducibility. The combination of descriptive statistics, correlation, and non-parametric testing provides a triangulated understanding of student achievement, question difficulty, and assessment effectiveness. This multi-method framework strengthens the validity of findings and supports informed decisions for curriculum development and instructional refinement.

It is important to restate here that this study aims to demonstrate how in-depth analysis of Quizizz-generated data can inform instructional improvement and promote continuous enhancement in language education. By leveraging the platform’s detailed reporting and applying rigorous statistical methods, educators gain valuable insights into student learning behaviors, question quality, and alignment with learning objectives. This data-driven process facilitates the identification of learning gaps and supports targeted interventions, ultimately contributing to more effective teaching strategies and improved student outcomes. The research highlights the potential of transforming raw assessment data into meaningful analytics that empower educators to make evidence-based decisions and foster sustainable improvements in language instruction.

RESULTS AND DISCUSSIONS

Due to space limitations, this paper presents only a carefully selected portion of the extensive data available from the Quizizz reports. The focus is primarily on student behavior and the development of the Mandarin 1 and Mandarin 2 courses. The results section highlights findings from four key assessments, which include the mid-term and final exams for each course. Descriptive statistics are used to summarize student performance. In addition, the results of Spearman correlation and Kruskal-Wallis tests are introduced to provide a basis for more detailed analysis. This focused presentation offers important insights into curriculum enhancement and instructional planning. It also sets the stage for a deeper examination of student learning outcomes and engagement patterns. The full dataset remains available for further research to support continuous course improvement.

Table 1. Descriptive Statistics of the Question Items Difficulty

No of Questions	20	25	16	30
Mean	67.35	65.08	67.63	66.00
Maximum	85	83	85	83
Minimum	21	42	29	39
Mode	67	69	74	73

Table 1 summarizes the descriptive statistics for question difficulty across four Mandarin language assessments. The average difficulty scores are tightly clustered between 65.08 and 67.63, indicating a consistently moderate level of challenge throughout the tests. This balance between easier and more difficult questions enables the assessments to effectively evaluate a broad range of student abilities. Such a distribution helps differentiate students with varying proficiency levels, which aligns with widely accepted educational assessment principles recommending a mix of question difficulties to improve both validity and reliability. The close similarity of mean difficulty scores across all four tests suggests that the assessments were carefully calibrated to maintain fairness and an appropriate level of challenge.

This descriptive overview works in line with other analyses, such as those in Table 2, to provide a comprehensive understanding of the assessment data. The consistency observed across these analyses reinforces the reliability of the findings and offers a strong foundation for evaluating student performance. Moreover, these results can guide instructional improvements by identifying areas where students may need additional support or where the curriculum might be adjusted to better meet learning objectives. Overall, the data demonstrate a thoughtful approach to assessment design that supports accurate measurement of student learning and informs ongoing curriculum development.

Table 2. Descriptive Statistics of Students' Test Scores

Mean	32.15	29.56	50.00	24.24
Max	38	37	60	30
Min	12	13	21	0
Mode	33	30	54	27
Accuracy	80.38	99.00	83.00	81.00

Table 1 summarizes the descriptive statistics for question difficulty across four Mandarin language assessments. The average difficulty scores are tightly clustered between 65.08 and 67.63, indicating a consistently moderate level of challenge throughout the tests. This balance between easier and more difficult questions enables the assessments to effectively evaluate a broad range of student abilities. Such a distribution helps differentiate students with varying proficiency levels, which aligns with widely accepted educational assessment principles recommending a mix of question difficulties to improve both validity and reliability. The close similarity of mean difficulty scores across all four tests suggests that

the assessments were carefully calibrated to maintain fairness and an appropriate level of challenge.

This descriptive overview works in tandem with other analyses, such as those in Table 2, to provide a comprehensive understanding of the assessment data. The consistency observed across these analyses reinforces the reliability of the findings and offers a strong foundation for evaluating student performance. Moreover, these results can guide instructional improvements by identifying areas where students may need additional support or where the curriculum might be adjusted to better meet learning objectives. Overall, the data demonstrate a thoughtful approach to assessment design that supports accurate measurement of student learning and informs ongoing curriculum development.

The analysis of question difficulty and student performance across four Mandarin language assessments reveals a well-balanced and effective evaluation design. Table 1 shows that the average question difficulty consistently falls within a moderate range, between 65.08 and 67.63, indicating a deliberate mix of easier and more challenging items. This balanced distribution allows the tests to accurately differentiate among students with varying proficiency levels, aligning with established best practices in educational assessment that emphasize validity and reliability through varied question difficulty. Complementing this, Table 2 highlights the variability in student scores, with mean scores ranging from 24.24 to 50.00, reflecting differences in accessibility and challenge across the assessments. The wider range of maximum and minimum scores further underscores the tests' ability to capture diverse student abilities. Together, these findings provide a comprehensive picture of the assessments' effectiveness, supporting targeted instructional improvements and curriculum refinement based on solid empirical data.

Table 3. Comparison of Students Level of Completion for the Four Tests

S. Cat	UTS M1	UAS M1	UTS M2	UAS M2
High (>80)	23 (85.19)	19 (70.37)	24 (82.76)	20 (68.97)
Moderate (60-79)	3 (11.11)	6 (22.22)	1 (3.45)	6 (20.69)
Low (<60)	1 (3.70)	2 (7.41)	4 (13.79)	3 (10.34)
Total	27	27	29	29

Table 3 shows a detailed comparison of student performance levels across the four Mandarin assessments by categorizing scores into high, moderate, and low groups. Most students achieved high scores on the mid-term exams, with 85.19% in UTS M1 and 82.76% in UTS M2 performing at this level. However, the final exams (UAS M1 and UAS M2) displayed a more balanced distribution of performance categories. Specifically, 70.37% and 68.97% of students scored highly in these finals, but there was also a noticeable increase in students falling into the low-performance group. This pattern suggests a need for formative assessments and targeted support to help lower-performing students. Black and Wiliam (1998) emphasize that such formative approaches can reduce achievement gaps by providing timely feedback and adapting instruction to meet individual needs.

Table 4 presents key statistical analyses examining student performance across the four Mandarin assessments. The Spearman correlation results show a moderate positive relationship between mid-term and final exam scores within each course. Specifically, the correlation coefficient for Mandarin 1 assessments is 0.4991 with a p-value of 0.00804, indicating a statistically significant association. For Mandarin 2, the correlation is slightly weaker at 0.37925 but remains significant with a p-value of 0.04246. These findings suggest that students who perform well on the mid-term tend to maintain similar performance on the final exam, supporting the reliability of the assessments and their ability to consistently measure related learning outcomes.

Table 4. Summary of Statistics Test

Spearman Correlation of UTS M1 and UAS M1
rs = 0.4991, p (2-tailed) = 0.00804.
The association between the two variables would be considered statistically significant.
Spearman Correlation of UTS M2 and UAS M2
rs = 0.37925, p (2-tailed) = 0.04246.
The association between the two variables would be considered statistically significant.
Kruskal-Wallis Test of UTS MI, UAS M1, UTS M2, UAS M2
The H statistic is 66.4526 (3, N = 96).
The p-value is < .00001. The result is significant at p < .05.

The Kruskal-Wallis's test reveals significant differences in student scores across all four assessments, with an H statistic of 66.4526 and a p-value less than 0.00001. This highly significant result indicates that at least one of the assessments differs in median performance compared to the others. Such variation may reflect differences in test difficulty, content, or question format. Importantly, the Kruskal-Wallis's test does not specify which assessments differ, only that differences exist. These results underscore the need for careful assessment design to ensure fairness and equal opportunity for students to demonstrate their knowledge. Together, the correlation and non-parametric analyses provide robust evidence of both consistency and variability in student performance, informing instructional refinement and curriculum development.

The findings of this study align well with established educational assessment theories, providing a theoretical framework to understand the observed patterns in question difficulty and student performance. The variation in question difficulty reflects the principles of Bloom's taxonomy, which emphasizes the importance of including questions that assess a range of cognitive skills, from foundational knowledge to higher-order thinking (Thamrin et al., 2019). The moderate difficulty levels and balanced distribution of easy, medium, and difficult questions indicate that the assessments were intentionally designed to challenge students at multiple cognitive levels. This approach ensures a comprehensive evaluation of learning, allowing educators to gauge not only basic recall but also students' abilities to analyze, evaluate, and create key stages in Bloom's revised taxonomy. By incorporating this range, the assessments can more effectively distinguish between different levels of student proficiency, supporting both validity and reliability in measurement.

Student performance variability observed in the data further supports the need for differentiated instruction; a concept grounded in Vygotsky's Zone of Proximal Development (ZPD) theory (Newman & Latifi, 2021). According to ZPD, learning is most effective when support is tailored to the learner's current abilities and scaffolded to help them progress to higher levels of understanding. The range of student outcomes suggests that while some learners have mastered the material, others require additional guidance and scaffolding. This highlights the importance of responsive teaching strategies that adapt to individual needs rather than adopting a one-size-fits-all approach. Differentiated instruction, therefore, becomes essential to ensure that all students can reach their potential, particularly in language learning contexts where proficiency levels often vary widely.

Patterns of student engagement, such as the number of attempts and distribution of scores, also reflect key principles of mastery learning. This educational approach advocates for providing students with multiple opportunities to demonstrate understanding and receive formative feedback that guides further learning (Ajjawi et al., 2022; Bardach et al., 2021). The data suggests that repeated attempts on assessments may contribute to improved performance, especially for students who initially struggle. This iterative process not only supports skill development but also fosters motivation and persistence, which are critical for language acquisition. Mastery learning thus promotes a growth mindset, encouraging students to view challenges as opportunities for improvement rather than fixed limitations.

The moderate positive correlations found between related assessments reinforce the concept of construct validity, which asserts that tests should accurately measure the intended learning outcomes (Brau, 2020). The statistically significant correlations between mid-term and final exam scores within each Mandarin course indicate that these assessments are consistent in evaluating similar constructs over time. This stability is crucial for educators to trust that assessment results meaningfully reflect student learning rather than random variation or measurement error. At the same time, significant differences in student performance across the four assessments, as revealed by the Kruskal-Wallis’s test, highlight the necessity of designing fair and well-aligned tests. According to educational measurement theory, assessments must provide equitable opportunities for all students to demonstrate their knowledge and skills (Constantinou & Wijnen-Meijer, 2022). Ensuring fairness involves careful consideration of content, difficulty, and format to avoid disadvantaging any group of learners.

This study highlights the critical role of thoughtful assessment design, formative feedback, and data-driven instructional adjustments in supporting diverse learners. The findings confirm the reliability and validity of the Mandarin language assessments while identifying areas where equity and accessibility can be enhanced. These insights have important implications for language education, emphasizing the need for ongoing evaluation and refinement of assessment practices to better meet the needs of all students. By integrating established educational theories with empirical data, educators can create more effective, inclusive learning environments that foster continuous growth and achievement.

CONCLUSION

This study demonstrates that analyzing Quizizz data from four Mandarin language assessments provides valuable insights into both question difficulty and student performance. The assessments consistently maintained a moderate difficulty level, featuring a variety of question types designed to evaluate a broad range of student abilities. Statistical analyses, including descriptive statistics and inferential tests, confirmed significant differences in student outcomes while supporting the reliability and validity of the assessments. While students generally performed well on mid-term exams, the final assessments proved more challenging, underscoring the importance of formative feedback and targeted instructional support to address learning gaps. The data also highlight the benefits of mastery learning approaches, where multiple attempts and ongoing feedback contribute to improved student achievement and engagement.

Educators are encouraged to leverage Quizizz analytics as a powerful tool to refine assessment design and tailor instruction, especially for students who face difficulties in final evaluations. Using a combination of descriptive and inferential statistical methods ensures a thorough and robust evaluation of learning outcomes, enabling educators to make informed decisions based on concrete evidence. This data-driven approach not only enhances student support but also promotes equity by identifying and addressing diverse learner needs. Furthermore, the integration of digital assessment platforms like Quizizz exemplifies how technology can transform language education, fostering more interactive, personalized, and effective learning experiences. As supported by recent studies (Yunus & Hua, 2021; Zulfa & Effendi, 2021), Quizizz creates an engaging environment that positively influences student motivation and achievement, making it an essential component of modern pedagogical practice.

REFERENCES

Ajjawi, R., Tai, J., & Dawson, P. (2022). Feedback for learning. In *International Encyclopedia of Education: Fourth Edition*. https://doi.org/10.1016/B978-0-12-818630-5.09013-8

Aulia, R., & Warni, S. (2024). Students' Perceptions Toward Quizizz as An Assessment Tool in EFL Classroom. *International Journal of Research in Education*, *4*(2), Article 2. https://doi.org/10.26877/ijre.v4i2.554

Bardach, L., Klassen, R. M., Durksen, T. L., Rushby, J. V., Bostwick, K. C. P., & Sheridan, L. (2021). The power of feedback and reflection: Testing an online scenario-based learning intervention for student teachers. *Computers and Education*, *169*. https://doi.org/10.1016/j.compedu.2021.104194

Brau, B. (2020). *Constructivism. The Students' Guide to Learning Design and Research* [Online post]. BYU Open Education. https://open.byu.edu/studentguide/constructivism

Constantinou, C., & Wijnen-Meijer, M. (2022). Student evaluations of teaching and the development of a comprehensive measure of teaching effectiveness for medical schools. *BMC Medical Education*, *22*(1). https://doi.org/10.1186/s12909-022-03148-6

Daulay, S. H., Ramadhan, A., & Wahyuni, S. (2023). Quizizz Application in Students Learning Outcomes: Teacher's and Students' Perception. *Modality Journal: International Journal of Linguistics and Literature*, *3*(2), 125. https://doi.org/10.30983/mj.v3i2.6719

Henukh, A., Made Astra, I., Supriyadi, Reski, A., & Hidayatullah, M. M. S. (2022). The Effectiveness of Using Quizizz in Fundamental Physics Learning in the Era of the Covid-19 Pandemic. *Journal of Physics: Conference Series*, *2309*(1), 012054. https://doi.org/10.1088/1742-6596/2309/1/012054

Mesterjon, Suwarni, Hermawansayah, Rulismi, D., Supama, Sahil, A., & Dali, Z. (2024). Effectiveness of the Use of Quizizz Media on Students' Learning Interest. *Futurity Education*, *4*(2), Article 2. https://doi.org/10.57125/FED.2024.06.25.13

Munawir, A., & Hasbi, N. P. (2021). THE EFFECT OF USING QUIZIZZ TO EFL STUDENTS' ENGAGEMENT AND LEARNING OUTCOME. *English Review: Journal of English Education*, *10*(1), 297–308. https://doi.org/10.25134/erjee.v10i1.5412

Newman, S., & Latifi, A. (2021). Vygotsky, education, and teacher education. *Journal of Education for Teaching*, *47*(1), 4–17. https://doi.org/10.1080/02607476.2020.1831375

Nurliana, E., & Nugroho, O. F. (2021). Analisis Hasil belajar dalam Penggunaan Quizizz Pada Pembelajaran IPA. *Seminar Nasional Ilmu Pendidikan Dan Multi Disiplin*, *4*(0). https://prosiding.esaunggul.ac.id/index.php/snip/article/view/139

Saepul, A. D., Helina, N., & Sutresna, Y. (2023). IMPROVING STUDENTS' LEARNING OUTCOMES THROUGH PJBL LEARNING MODELS IN PRACTICES FOR MAKING OF CASTING TAPE (Manihot utilissima) WITH THE ASSISTANCE OF MEDIA QUIZIZ. *Journal Of Biology Education Research (JBER)*, *4*(1), 25–30. https://doi.org/10.55215/jber.v4i1.7583

Thamrin, N. R., Widodo, P., & Margana. (2019). Developing Higher Order Thinking Skills (Hots) for Reading Comprehension Enhancement. *Journal of Physics: Conference Series*, *1179*(1). https://doi.org/10.1088/1742-6596/1179/1/012073

Umam, M. A. K., Sukmanasa, E., Heldayanti, H., & Pratiwi, I. E. (2025). Enhancing Student Learning Outcomes Through Quizizz-Supported Culturally Responsive Teaching in Problem-Based Learning. *Journal of General Education and Humanities*, *4*(2), 351–360. https://doi.org/10.58421/gehu.v4i2.402

Yunus, C. C. A., & Hua, T. K. (2021). Exploring a Gamified Learning Tool in the ESL Classroom: The Case of Quizizz. *Journal of Education and E-Learning Research, 8*(1), 103–108. https://doi.org/10.20448/journal.509.2021.81.103.108

Zhao, F. (2019). Using Quizizz to Integrate Fun Multiplayer Activity in the Accounting Classroom. *International Journal of Higher Education*, *8*(1), 37. https://doi.org/10.5430/ijhe.v8n1p37

Zulfa, S., & Effendi, A. (2021). Teacher's Illocutionary Acts in Online Learning Interactions. *Jurnal Pendidikan Dan Pengajaran*, *54*(1). https://doi.org/10.23887/jpp.v54i1.33061