MERAL Publications and Solutions Pvt. Ltd.

Original Article

Evaluating ZipGrade for automated multiple-choice scoring: a pilot study at a Vietnamese medical university

Ho Nguyen Anh Tuan[1], Cao Nguyen Hoai Thuong[1], Tran Tien Tai[1], Nguyen Dung Tuan[1], Nguyen Thanh Van[2]*

[1]Faculty of Medicine, Pham Ngoc Thach University of Medicine, Ho Chi Minh city, Vietnam. [2]School of Medicine and Pharmacy, Tra Vinh University, Tra Vinh province, Vietnam.

Correspondence: Nguyen Thanh Van, School of Medicine and Pharmacy, Tra Vinh University, Tra Vinh province, Vietnam. drthanhvan@gmail.com

Received: 02 December 2024; **Revised:** 16 March 2025; **Accepted:** 17 March 2025

ABSTRACT

The study was conducted to determine the validity and reliability of ZipGrade software in marking multiple-choice exams and compare it with conventional manual marking methods. Implementation time is from October 2023 to January 2024 at Pham Ngoc Thach University of Medicine. Validity is determined by the proportion of tests scored with ZipGrade software whose results completely match the standard results out of the total number of tests, and reliability is determined by the intraclass correlation coefficient (ICC) between two scores of each method and between each method and the standard result. After performing multiple-choice testing on 180 tests, research shows that ZipGrade software has high accuracy (93 to 100%) and high reliability (ICC coefficient from 0.982 to 0.999). Compared with the manual method, ZipGrade software also has statistically significantly higher accuracy and reliability ($p < 0.001$). The average time to grade a multiple-choice test using the manual method is 6 to 7.7 times that of ZipGrade software. Schools should use ZipGrade software to grade multiple-choice exams and organize training sessions to guide teachers on using it.

Keywords: Multiple-choice question, Reliability, Scoring validity, ZipGrade

Introduction

In the teaching activities of teachers and lecturers at educational institutions, grading exams is one of the most time-consuming tasks, requiring both patience and high accuracy. Currently, multiple-choice testing is widely applied in most schools across the country due to its numerous advantages over other testing methods. With a large volume of tests and the associated pressure of meeting regulatory deadlines for publishing scores to students within a specific timeframe after the examination, finding a quick, effective, and automated grading solution is crucial [1]. Recognizing the needs of the majority of educators, various software tools have been developed to support teaching activities, including multiple-choice grading software, which has already shown significant promise and proven its effectiveness [2, 3].

Among these tools, ZipGrade, a mobile application for grading multiple-choice exams, is widely used due to its simplicity, convenience, and efficiency. Besides its listed advantages, ZipGrade is also known as a licensed software available for free with a limit of 100 graded tests per month. Users who wish to access unlimited grading can pay a fee of $7 (approximately 160,000 VND) per year—a relatively low cost compared to the benefits it offers. In one experiment conducted at an English language center using ZipGrade to grade over 400 multiple-choice answers from the TOEIC Bridge test, the results were processed in just 2-3 seconds, demonstrating the application's speed [2, 4, 5]. In addition to its prominent features, ZipGrade also offers score storage and export capabilities, which are highly

Access this article online	
Website: www.japer.in	**E-ISSN:** 2249-3379

How to cite this article: Tuan HNA, Thuong CNH, Tai TT, Tuan ND, Van NT. Evaluating ZipGrade for automated multiple-choice scoring: a pilot study at a Vietnamese medical university. J Adv Pharm Educ Res. 2025;15(2):95-9. https://doi.org/10.51847/8PjaeAguEN

convenient for notifying students of their scores and managing classes. Moreover, ZipGrade can analyze test results, providing insights into the percentage of correct and incorrect answers for each question. Based on these analyses, teachers can identify questions with high error rates and use this information to reteach unclear concepts or apply appropriate improvement measures.

At Pham Ngoc Thach University of Medicine, multiple-choice testing is predominantly used for final exams, with test questions displayed on a computer and students answering by shading their responses on a pre-designed answer sheet. With an annual enrollment of approximately eight thousand students across all programs, the number of exams per testing period is enormous. This situation not only requires the mobilization of a large number of lecturers for grading but also imposes significant time and labor pressure on exam organization and management. To automate the grading process and reduce the workload for lecturers, the research team piloted the application of the ZipGrade software for multiple-choice test grading at the university [6, 7]. This study was conducted to evaluate the validity and reliability of the ZipGrade software during its pilot implementation at Pham Ngoc Thach University of Medicine and to compare its performance with conventional manual grading methods.

Materials and Methods

Study design

This study employed a cross-sectional design and was conducted at Pham Ngoc Thach University of Medicine between October 2023 and January 2024. The focus was on evaluating the performance of the ZipGrade software in grading multiple-choice exams.

Study subjects and sample size

The subjects of the study included the ZipGrade software and multiple-choice exams. Two answer sheet formats were used: the default manual answer sheet and a custom-designed answer sheet created in the ZipGrade application.

The sample size was determined using the formula for estimating agreement based on the intraclass correlation coefficient (ICC), where α: Type I error probability, with $\alpha = 0.05$, leading to $Z_{(1-\alpha/2)} = 1.96$; k: Number of raters. For each grading method (ZipGrade and manual), two independent raters graded the exams, and one summarized and established the reference results, resulting in $k = 5$; ρ: ICC coefficient. Based on the absence of prior studies, the study used $\rho = 0.9$; d: Margin of error for the ICC, set at $d = 0.05$. Using these parameters, the required sample size was calculated as 34 exams. To enhance robustness, the study graded 80 exams using the default answer sheet and 100 exams using the custom-designed answer sheet, resulting in a total of 180 graded exams.

$$n \geq 1 + \frac{2Z_{1-\alpha/2}^2(1-\varrho)^2[1+(k-1)\varrho]^2}{k(k-1)d^2} \quad (1)$$

Sampling technique

Convenient sampling was employed in three steps: Step 1: An examination session with approximately 100 students was randomly selected. The Anatomy exam for postgraduate students was chosen. Step 2: Students were randomly divided into two groups. One group used the default answer sheet, and the other used a custom-designed answer sheet. The exam, which was displayed on a computer, consisted of 100 multiple-choice questions and lasted 90 minutes. Step 3: After the examination, the research team collected, counted, and sealed the exams per institutional regulations. The final sample included 80 exams graded using the default answer sheet and 100 exams graded using the custom-designed sheet.

Study variables

The study evaluated four primary variables: validity, reliability, grading time, and average grading time per exam. Validity was defined as the proportion of exams graded by ZipGrade that matched the reference results out of the total number of exams. Reliability was measured using the intraclass correlation coefficient (ICC), assessing consistency between the two grading attempts of each method (ZipGrade or manual) and between each method and the reference results. Grading time was calculated as the total time required for all grading steps, which included answer key creation, grading, and error handling for ZipGrade, sorting exams by test codes, creating answer keys, grading, and entering scores for the manual method. The average grading time per exam was determined by dividing the total grading time for both attempts by the number of exams.

Data collection

The study involved grading multiple-choice exams using both the manual method and the ZipGrade software. Each exam contained 100 multiple-choice questions with four answer options (A, B, C, D) and a 90-minute duration. Students were randomly divided into two groups, with one group using the default answer sheet and the other using a custom-designed sheet. Grading was performed as follows: In ZipGrade, each exam was graded twice (first and second attempts). In the manual method, each exam was manually graded twice (third and fourth attempts). Manual grading results were entered into Microsoft Excel, while ZipGrade results were directly exported to an Excel file. The research team consolidated all results. Discrepancies in the number of correct answers across the four grading attempts were manually rechecked to establish the reference results.

Data analysis

Data were entered and managed using Microsoft Excel and analyzed with Stata 14.0. Validity was assessed by comparing each method's results with the reference results. Reliability was

evaluated using ICC to measure consistency within each method and between each method and the reference results. A one-sample binomial test was used to compare the validity of the ZipGrade software and the manual grading method.

Ethical considerations

The study ensured that no harm or adverse effects were caused to the work, examinations, or activities of students, faculty, or university staff. All processes adhered to ethical standards, protecting the confidentiality and integrity of the participants and the institution.

Results and Discussion

Table 1. Validity of the ZipGrade software

ZipGrade	Default answer sheet (n = 80)		Custom-designed answer sheet (n = 100)	
	n	%	n	%
Validity				
First attempt	80	100	95	95
Second attempt	78	97,5	93	93

For the default answer sheet, the first attempt achieved absolute accuracy (100%), while the second attempt showed 97.5% accuracy. On the custom-designed answer sheet, the accuracy was slightly lower, with 95% on the first attempt and 93% on the second attempt.

Table 2. Reliability of the ZipGrade software

ZipGrade	Correct Answers (Mean ± SD)		ICC
	First attempt	Second attempt	
Default Answer Sheet (n = 80)	21,5 ± 5,62	21,5 ± 5,61	0,999
Custom-Designed Answer Sheet (n = 100)	24,2 ± 4,2	24,1 ± 4,2	0,982

When assessing consistency across two attempts, the default answer sheet demonstrated higher reliability, with an ICC of 0.999. Nonetheless, the custom-designed answer sheet also displayed excellent consistency, with an ICC of 0.982.

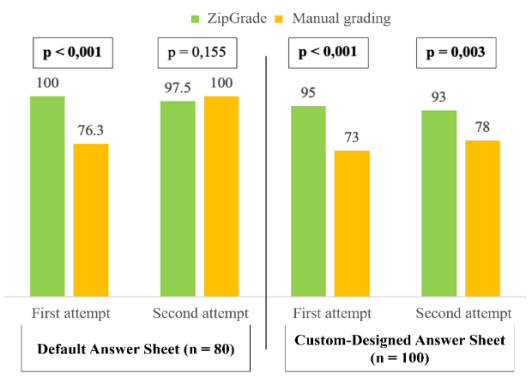

Figure 1. Comparison of Validity Between ZipGrade Software and Manual Grading

When comparing ZipGrade and manual grading, the ZipGrade software demonstrated significantly higher accuracy in the first attempt with the default answer sheet and both attempts with the custom-designed answer sheet. The only instance where manual grading showed higher accuracy was in the second attempt with the default answer sheet, but the difference was not statistically significant.

Table 3. Comparison of Reliability Between ZipGrade Software and Manual Grading

	$ICC_{first\ attempt}$	$ICC_{second\ attempt}$	$ICC_{overall}$
Default Answer Sheet (n = 80)			
ZipGrade Software	1	0,9996	0,9997
Manual Grading	0,996	1	0,997
Custom-Designed Answer Sheet (n = 100)			
ZipGrade Software	0,998	0,979	0,986
Manual Grading	0,979	0,985	0,978

The alignment of ZipGrade results with the reference standard was highly consistent, with an overall ICC of 0.9997 for the default answer sheet and 0.986 for the custom-designed answer sheet. When comparing the two methods against the reference standard, the ZipGrade software demonstrated higher reliability than the manual grading method for both types of answer sheets.

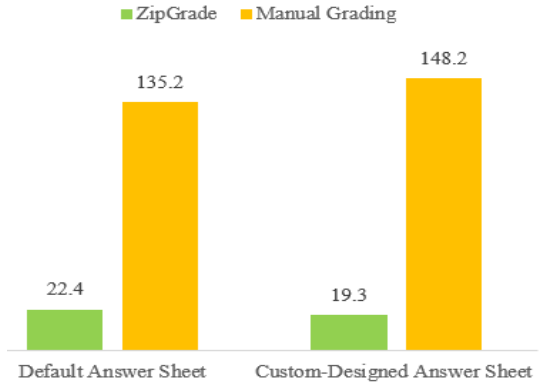

Figure 2. The average time to grade an exam using ZipGrade software and the manual grading method (in seconds)

When comparing the average time required to grade an exam using the two methods, the ZipGrade software demonstrated a significantly faster grading time. Manual grading took 6 to 7.7 times longer than ZipGrade.

Validity and reliability of zipgrade software in multiple-choice exam grading

Several studies have been conducted to evaluate the use of ZipGrade software for student assessment. However, most of these studies focus on user experiences, practical applications, and faculty acceptance of the software. Few, if any, have assessed its validity and reliability as comprehensively as our study.

Ningsih and Mulyono (2019) examined how teachers used and perceived digital assessment tools such as "Kahoot!" and ZipGrade in classrooms. The survey results revealed positive

attitudes among teachers toward the adoption of these applications. Additionally, the study highlighted factors that encouraged teachers to use these tools, including creating a fun learning environment, practicality, automated scoring, and instant feedback [8]. Similarly, Suhendara *et al*. (2020) reported that teachers had positive experiences using ZipGrade for analyzing student performance. The software significantly supported teachers in efficiently correcting student answers. Students also found ZipGrade's answer sheets easy to use, appreciated the ability to view exam results immediately after leaving the exam room, and valued having their answer sheets photographed for review. Teachers noted that the software streamlined their workflow, particularly in analyzing assessment results [9, 10].

In early 2023, a study investigated the impact of Web 2.0 tools on students' year-end academic performance. The results showed that students exposed to applications like ZipGrade and Padlet scored higher on exams. This study suggested incorporating such applications into regular curricula [11].

Comparison of validity, reliability, and grading time between zipgrade and manual grading

The study revealed that ZipGrade exhibited significantly higher accuracy compared to manual grading. Similar findings were observed regarding reliability. These results align with recent studies, such as one conducted in Iraq in 2023 [11], and another in the Philippines in 2019 [12, 13]. The Iraq study found that compared to traditional methods, ZipGrade was more cost-effective, demonstrating 100% accuracy versus 94% for manual grading. Additionally, ZipGrade required only 3 seconds per test, whereas manual grading took 58 minutes [12]. Another study from 2019 conducted on teachers at Ternate West National High School in the Philippines highlighted that ZipGrade improved the timeliness of teacher report submissions. Using ZipGrade reduced the average reporting time to 3.4 days compared to 6 days with traditional methods [14].

In the context of Pham Ngoc Thach University of Medicine, using ZipGrade offers significant practical benefits, including saving substantial time and effort for instructors while providing detailed analysis of exam results. Although ZipGrade is licensed software, its cost - approximately 170,000 VND per year per account is highly cost-effective. Moreover, its superior accuracy and reliability compared to manual grading make it a valuable tool for multiple-choice exam grading.

Limitations

Despite the promising findings of this study, some limitations should be acknowledged. First, the study was conducted within a single medical university in Vietnam, which may limit the generalizability of the results to other academic institutions with varying exam formats, curricula, or student populations. Second, the scope of this study focused on the validity and reliability of ZipGrade in grading multiple-choice exams; however, it did not explore the software's integration into broader assessment systems or its long-term impact on teaching and learning outcomes. Finally, the study did not examine technical challenges, such as hardware or software compatibility issues, which could influence the practical implementation of ZipGrade in diverse educational settings.

Recommendations

To enhance the utility and applicability of ZipGrade, future studies should explore its performance in a broader range of educational contexts, including different universities and disciplines. Additionally, researchers should investigate the software's effectiveness when used in combination with other digital assessment tools to provide a more holistic evaluation framework. It is also recommended that training workshops for faculty members emphasize not only the operational use of the software but also strategies for integrating it into pedagogical practices to maximize its educational benefits. Lastly, assessing the cost-effectiveness of ZipGrade over an extended period can provide further insights into its value for institutions with varying budget constraints.

Conclusion

ZipGrade software has been proven to exhibit high validity and reliability while significantly reducing grading time compared to traditional methods. It is recommended that the university adopt ZipGrade for grading multiple-choice exams and organize training sessions to help instructors use the software effectively.

Acknowledgments: The authors would like to thank the leaders of Pham Ngoc Thach University of Medicine for their participation and support during the study.

Conflict of interest: None

Financial support: None

Ethics statement: The study was reviewed and approved by the Ethics Committee of Pham Ngoc Thach University of Medicine. No identifiable information was collected. All data used in the study were anonymized and securely stored in accordance with the institutional regulations.

References

1. Cherner T, Lee CY, Fegely A, Santaniello L. A detailed rubric for assessing the quality of teacher resource apps. J Inf Technol Educ: Innov Pract. 2016;15:117-43.
2. Stephen MP. ZipGrade: scan response forms with your phone. The Language Teacher. 2018:29-32.
3. Choudhary V, Sharma S, Vashishtha S, Malik A. Recent findings, application and future direction of natural

extracts: mucilage. Int J Pharm Phytopharmacol Res. 2023;13(1):33-43. doi:10.51847/EAUqALnIHP

4. Uzun K, Karataş Z. Investigating the role of metacognitive beliefs, ambiguity tolerance, and emotion processing in predicting nurses' generalized anxiety disorder. J Integr Nurs Palliat Care. 2022;3:36-42. doi:10.51847/mXbCbDAVpU
5. Al-Twaijri SA, AlKharboush GH, Alohali MA, Arab IF, Alqarni RH, Alharbi MS. Application of lasers for soft tissues in orthodontic treatment: a narrative review. Bull Pioneer Res Med Clin Sci. 2024;3(1):1-6.
6. Alqahatani S. Study on machine learning and deep learning in medical imaging emphasizes MRI: a systematic literature review. Int J Pharm Res Allied Sci. 2023;12(2):70-8. doi:10.51847/kj4hoW5tIZ
7. Aburas M. Characterization and identification of pantoea calida from contaminated soil and its biocontrol by streptomyces coeruleorubidus. World J Environ Biosci. 2022;11(3):50-6. doi:10.51847/JxVBFGWSn1
8. Ningsih SK, Mulyono H. Digital assessment resources in primary and secondary school classrooms: teachers' use and perceptions. Int J Interact Mob Technol. 2019;13(8):167. doi:10.3991/ijim.v13i08.10730
9. Suhendara, Surjonob HD, Slamet PH, Priyanto. The effectiveness of the ZipGrade-assisted learning outcomes assessment analysis in promoting Indonesian vocational teachers' competence. Int J Innov Creat Change. 2020;11(5):701-19.
10. Gioia G, Freeman J, Sipka A, Santisteban C, Wieland M, Gallardo VA, et al. Study of bacterial contamination of house flies in different environments. Entomol Appl Sci Lett. 2023;10(4):56-61. doi:10.51847/Rb6CEz672N
11. Kara S. The effects of web 2.0 tools on foundation english students` success rates at a private university in Iraq. Int J Soc Sci Educ Stud. 2023;10(1):22-36.
12. Palanas DM, Alinsod AA, Capunitan PM. Digital assessment: empowering 21st century teachers in analyzing student's performance in calamba city. JPAIR Instit Res. 2019;13(1):1-4. doi:10.7719/irj.v13i1.779
13. Joseph KA, Ahuja S, Zaheer S. Secondary ovarian malignancy in an imatinib treated chronic myeloid leukemia patient diagnosed on fluid cytology. Clin Cancer Investig J. 2023;12(4):10-3. doi:10.51847/y7ma3YBrbY
14. Cimafranca JRG, Tamayo L. Using ZipGrade application in TLE teachers' reports compliance. AAJMRA. 2019;3(2).