Conducting a Trial of Automated Open-ended Question Scoring in Mathematics Classes for Diagnosing Mathematical Proficiency through a Digital Learning Platform

Jiraprapa Chaiyawut[1], Apinya Foithong[2], Putcharee Junpeng[3], Samruan Chinjunthuk[4], Prapawadee Suwannatrai[5], Mongkhol Prasertsang[6], Metta Marwiang[7]
[1,2,3,4,5,6,7]Faculty of Education, KhonKaen University, 40002 KhonKaen, Thailand
[1]jichai@kku.ac.th, [2]apinya_ann@kkumail.com, [3]jputcha@kku.ac.th,
[4]samchi@kku.ac.th, [5]prasu@kku.ac.th, [6]mongpra@kku.ac.th, [7]metmaw@kku.ac.th

Abstract

The primary aim of the current study was to analyze the effectiveness of the automated open-ended scoring in providing insights into the assistance of students in terms of improving their mathematical proficiency levels through a real-time digital learning platform. In the preliminary study, the research team has developed a web application system that is a measurement tool for students to self-regulate their learning progress for automated feedback to diagnose their mathematical proficiency levels. This paper reports the results of a trial on automated open-ended question scoring in mathematics classes as a continuous phase from the preliminary study. A total of 106 seventh-grade students were randomly selected from a population of 528 of them from 23 schools located in the northeastern region of Thailand. Then, the researchers used a stratified random sampling method to select four mathematics classes from four different sizes of schools, namely small, medium, large, and extra-large. The researchers employed design-based research along with three phases and used a survey research design for the trial on the automated open-ended question scoring system. The results indicated that the trial of the automated open-ended question scoring justifying the majority of seventh-grade students' mathematical proficiency levels was consistent with their actual grades, based on the interview results with their teachers. Moreover, the test results were imported into the created web application system consisting of four sections, namely import data, processing, display, and assessment report sections. Then, the researchers focused on the effectiveness of the automated open-ended question scoring system and whether it is fit the real setting through the nine experts' heuristic evaluation. The heuristic results indicated that the automated open-ended scoring system of the digital learning platform was found effective in terms of its usefulness, interpretation, and correctness at the most appropriate levels from the nine experts' perspectives. Finally, the results showed that there were positive, strong, and significant correlations between the mathematical process and conceptual structural dimensions in terms of mathematical proficiencies, [r(106) = 0.782] for the subject of Measurement and Geometry [r(106) = 0.815] with seventh-grade students' examination achievement at a significant level of 0.01. In conclusion, the real-

*e*ISSN1303-5150

time digital learning platform and the open-ended question scoring system itself were found effective to diagnose students' mathematical proficiency levels.

KeywordsAutomated open-ended question scoring, digital learning platform, mathematical proficiency level, seventh-grade students.

DOI Number: 10.48047/NQ.2022.20.16.NQ880380 NeuroQuantology2022;20(16):3759-3768

INTRODUCTION

Mathematics is a fundamental discipline and crucially important in terms of motivating students' knowledge development [1]. [1] emphasized the importance of mathematics as a method of real-world status for numerous scientific disciplines. This was supported by [2] who indicated that mathematical proficiencies should be the 21st Century skills needed to resolve the complications of daily life, both proficiently and suitably. [3] emphasized the importance of a digital learning platform as an integral part of teaching and learning mathematics because it can be used as a measurement tool to diagnose the students' mathematical proficiency levels, and how they can be addressed to enhance learning.

Online learning has grown tremendously, particularly during the coronavirus 2019 (COVID-19) pandemic because it is more flexible than the traditional educational environment, as highlighted by [4]. Therefore, a digital learning platform for the purpose of facilitating online learning such as an automated open-ended question for diagnosing students' mathematical proficiency levels has emerged in the past decades [5]. Mathematical proficiency is defined as a student's capability to search, speculate, and think logically in the cognitive process to comprehend how to solve a mathematical problem by using appropriate strategies to solve problems and replicate the procedure used to solve the problems [6].

Item Response Theory has various advantages in terms of sample independent parameters and estimates, increase precision in estimates, standards metrics of estimates between subjects and between tests, the simpler composition of tests for specific audiences, for example, item bank, and computerized adaptive testing [7]. According to [7], Multidimension Test Response Theory means mathematics items with two factors, namely arithmetic problem solving and algebraic symbol manipulation. Therefore, the application of Multidimension Test Response Theory allows a single item to provide information for more than one dimension, test length can be reduced significantly. Moreover, Multidimension Test Response Theory can identify the extent to which the underlying elements are causing unexpected invariance. Finally, Multidimension Test Response Theory progresses in field such as abnormal response patterns.

RESEARCH PURPOSES

The primary purpose of this research was to develop an automated open-ended question scoring to assess seventh-grade students' mathematical proficiency level through a real-time digital learning platform. Specifically, the researchers intended to analyze students' response patterns, compare the scoring methods, and assess the quality of an open-ended question examination between dichotomous scoring and polytomous scoring. This was followed by developing an automated open-ended question scoring and examining its quality. Finally, the researchers conducted a trial using the created automated open-ended question scoring system in four mathematics classes. However, this paper will report on the final purpose only.

*e*ISSN1303-5150

MATERIALS AND METHODS

Research Design and Samples

The researchers used a design-based research methodology consisting of three phases in this research. The design-based research methodology is found suitable for this research because the basic process of this methodology involves developing solutions, so-called interventions, to the research problems. Then, the interventions are used to test how well they work [8, 9].

A survey research design utilizing mixed-mode methods was employed. The mixed-mode methods consisted of both quantitative and qualitative approaches used to achieve the research purposes as stated above. Mixed-mode methods were suitable for this research because researchers could obtain a more substantial indication of the practicality of the automated open-ended question scoring of the real-time digital learning platform which functions as a digital mathematical assessment technology system than a standalone quantitative or qualitative method [10]. The quantitative approach uses multiple survey questions about automated open-ended question scoring of the real-time digital learning platform to collect data from 106 seventh-grade students and nine experts comprised of two mathematics teachers who are teaching seventh-grade mathematics, two computer engineers, three experts in the field of educational measurement and evaluation, and two experts in the field of technology.

The population of seventh-grade students and schools located in the northern region of Thailand is 528 and 23 respectively in the academic year 2022. A total of 106 seventh-grade students were nominated from a population of 528 using a stratified random sampling technique according to their school size as test-takers. This permits researchers to acquire a sample population including small, medium, large, and extra-large school sizes that greatest symbolize the whole population being studied [11]. The four research schools have sufficient computers and Internet access for students to take the examination. Table 1 shows the distribution of the test-takers in the four different types of school sizes.

Table 1. Distribution of the test-takers

No.	School	School Size	No. of Test-takers
1.	S1	small	11
2.	S2	medium	20
3.	S3	Large	29
4.	S4	Extra-large	45
	Total		106

Furthermore, the researchers also steered in-depth interviews with the nine experts in the condition where researchers planned to explore the practicality of the automated open-ended question scoring of the real-time digital learning platform from the experts' professional views. On top of that, the nine experts also contributed to a heuristic evaluation in order to offer remarks about automated open-ended question scoring of the real-time digital learning platform quantitatively and qualitatively.

Research Procedure

In the first phase, the researchers explored the mathematical proficiencies of the seventh-grade students according to their test results to design a prototype of a mathematical proficiencies assessment tool through a real-time automatic digital platform to report their machine learning. Hence, the researchers generated the progress maps in each mathematical proficiencies dimension to fit the real setting.

The preliminary results of the first phase showed that there are five levels of mathematical proficiency for each dimension. The progress map of the mathematical process dimension consisted of non-response/irrelevance, unrecalled memory, basic memory and reproduction, simple skills and concept, and strategic or

extended thinking while the conceptual structure dimension has five levels as well, namely non-response, pre-structure, uni-structure, multi-structure or relation structure, and extended abstract structure. Figure 1 illustrates progress maps that are used by researchers as a precise guide to measuring the two dimensions of seventh-grade students' mathematical proficiency levels.

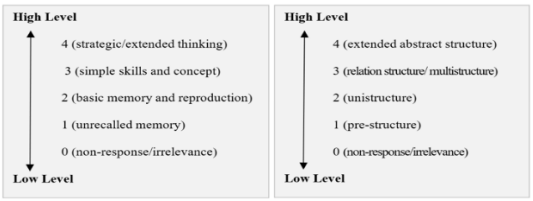

Figure 1. Progress maps of the mathematical process dimension and the conceptual structure dimension

In the second phase, the researchers used multidimensional random coefficients multinomial logit model to verify the created mathematical proficiency measurement model. The results of the second phase have proved that it is compliant with the quality of the real-time digital learning platform

In the final phase, the researchers conducted a trial using the created automated open-ended question scoring system in four mathematics classes involving 106 seventh-grade students. Moreover, a group of nine experts participated in a heuristic evaluation and an overall system evaluation of the automated open-ended question scoring system of the digital learning platform was performed quantitatively. Besides, the researchers continued to revise the heuristic evaluation ideas of [12] to evaluate the quality of the automated open-ended scoring system of the digital learning platform for diagnosing the levels of mathematical proficiency of seventh-grade students in three facets, namely usefulness, interpretation, and correctness.

A five-point Likert scale assessment instrument was used as a tool to evaluate the experts' interpretations ranging from minimal to the most appropriate levels for the automated open-ended question scoring system. Table 2 presents the five appropriate levels which were classified in accordance with the mean score range.

Table 2. Classification of appropriate levels according to the mean score range

Mean Score Range	Interpretation
1.00 to 1.49	Minimal appropriate
1.50 to 2.49	Less appropriate
2.50 to 3.49	Medium appropriate
3.50 to 4.49	Very appropriate
4.50 to 5.00	Most appropriate

Data Analysis

The researchers started to analyze the examination scores obtained from 106 seventh-grade students as users as they upload their answers to the digital learning platform in order to investigate the effectiveness of the system in terms of its accuracy while comparing it to actual examination scores. The trial of the automated open-ended question scoring was analyzed using descriptive statistics to summarize and organize the characteristics of the test results. This was followed by using the Pearson correlation to examine the linear relationship between the two mathematical proficiency dimensions and the total scores of the 106 students' examination achievement.

A total of nine experts' evaluation was employed as one of the methods to verify the practicality of the automated open-ended question scoring system used in this research. The experts' evaluation could assist researchers to classify and assign an assessment of the effectiveness of the system that could not be quantified. Once again, the researchers utilized mean scores and standard deviations to analyze the heuristic evaluation assessed by the nine experts. Finally, the researchers interviewed the nine experts after they completed their heuristic evaluation. The interview data were analyzed using content analysis.

*e*ISSN1303-5150

RESULTS

Results of Experts' Heuristic Evaluation According to Students' Automated Open-ended Question Scoring

The effectiveness of the automated open-ended question scoring system of its usefulness, interpretation, and correctness was measured based on the automated open-ended question scoring participated by 106 seventh-grade students in the digital learning platform. Table 3 shows the heuristic evaluation results according to the 106 seventh-grade students' open-ended questions scoring reports. The results indicated that the automated open-ended question scoring system of the digital learning platform was found effective in terms of its usefulness, interpretation, and correctness. Precisely, the highest mean score was the usefulness of the system (mean score = 4.89, *SD* = 0.58) which was found at the most appropriate level. This implies that the automated open-ended question scoring system can be used to diagnose their mathematical proficiency levels appropriately besides matching users' needs. Therefore, there was no doubt that the feedback reflection methods can be used in a concrete manner.

On top of that, the interpretation and correctness features of the automated open-ended question scoring system of the digital learning platforms were found at the most appropriate level with the respective mean score of 4.83 (*SD* = 0.29) and 4.78 (*SD* = 0.58). The interpretation of the automated open-ended question scoring system shows the description and scope of reporting feedback content are appropriately interpreted. Therefore, it can be concluded the automated open-ended question scoring design is considered comprehensive and consistent with current assessment guidelines. Moreover, the assessment objectives and evaluation process of the automated open-ended question scoring system are correctly processed. Therefore, the automated open-ended question scoring system can correctly designate students' actual mathematical proficiency levels. All three quality features of the automated open-ended system are found at the most appropriate levels from the experts' perspectives.

Table 3. Heuristic evaluation results in the effectiveness of an automated open-ended question scoring system

No.	Assessment Item	Mean Score	*SD*	Interpretation
Usefulness				
1	Scoring formats have matched the needs of users.	5.00	0.00	Most appropriate
2	Scoring reflection reports can be utilized in a concrete manner.	4.67	0.58	Most appropriate
3	Scoring methods can be utilized to enhance students' mathematical proficiency levels.	5.00	0.00	Most appropriate
Overall usefulness		4.89	0.58	Most appropriate
Interpretation				
4.	The description of	4.67	0.58	Most appropriate

	automated open-ended question scoring feedback reports is clearly indicated.			
5.	The scope of reporting scoring content is appropriate.	5.00	0.00	Most appropriate
6.	The scoring report method is comprehensive.	5.00	0.00	Most appropriate
7.	The feedback reporting system is consistent with current assessment guidelines.	4.67	0.58	Most appropriate
Overall interpretation		4.83	0.29	Most appropriate
Correctness				
8.	Assessment objectives are correctly indicated.	5.00	0.00	Most appropriate
9.	The evaluation process of the system is correct.	5.00	0.00	Most appropriate
10.	Scoring reports are correctly indicated students' actual mathematical proficiency levels.	4.33	0.58	Very appropriate
Overall accuracy		4.78	0.58	Most appropriate
Overall effectiveness		4.83	0.23	Most appropriate

Experts' Interview Results Regarding the Open-ended Scoring System

The researchers conducted in-depth interviews with the nine experts after they finished the heuristics evaluation. The following excerpts are the summary of the experts' opinions to enhance the automated open-ended question scoring system.

"*This automated open-ended question scoring system seems to be concrete and can observe the implementation process in the real context. These benefits students, teachers, and those involved as users. However, the data preparation process still has to be improved. This is because the system may be expanded to many other groups in the future*," said Expert 1.

"I found the automated open-ended question scoring system is very useful and very practical as well. However, if there is a comprehensive development without a data preparation process, the system will be more substantial. Anyway, I found this system is very useful to the students," mentioned Expert 5.

Based on the above interview results, researchers concluded that the scoring design for the test via digital technology has to improve at the first section, that is import section whereby it is the data preparation section as shown in Figure 2.

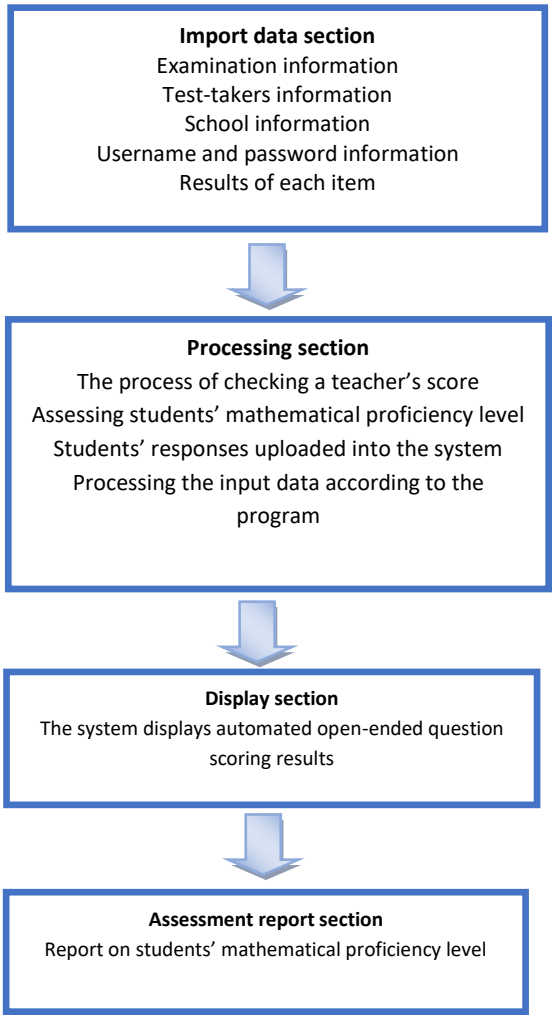

Figure 2. Automated open-ended question scoring design via digital technology

The qualitative results of the differences between students' actual results and automated open-ended scores, the following excerpts from the nine experts could help researchers to make a conclusion about there is no difference between the actual test results and the automated scores as follows.

"*The design of the scoring system is easy to use and provides students' examination results without teachers having to waste time checking and computing the scores. Based on the automated open-ended question scoring results compared to the real scores and their mathematical proficiencies, they are very close and consistent as much as possible*," mentioned Expert 2.

"The results from the automated scoring system, it can be seen that each student's mathematical proficiency level and his or her actual test scores are oriented in the same direction. The screen in the assessment report section is easy to understand and can get an overview of the overall operating system with a clear procedure," highlighted by Expert 3

C. Results of Experts' Evaluation of the Effectiveness of Automated Open-ended Question Scoring System

The effectiveness evaluation results were reported by the nine experts regarding the quality of the automated open-ended question scoring system of the digital learning platform indicating that all the features are at the most appropriate levels. This means that they came to absolutely agree with the appropriateness of the system in the following features of the automated open-ended question scoring system in terms of (i) visibility of system status; (ii) match between the system and the real world; (iii) user control and freedom; (iv) flexibility and efficiency of use; (v) system help and manuals, and (vi) support and enhance user skills (mean score = 5.00, *SD* = 0.00).

Nevertheless, the other features of the automated feedback system still reached the most appropriate levels, the mean score of 4.67. The features include (i) consistency and standards; (ii) error prevention; (iii) memory recall and retrieval system; (iv) help users recognize, diagnose, and recover from errors; (v) satisfaction and acceptance of user interactions, and (vi) protection of personal information with the mean score as 4.67. However, the system evaluation showed that there is only one feature of the automated feedback system that needs to

be improved, namely beautiful and simple design (mean score = 4.33, *SD* = 0.57). Generally, the automated feedback system of the digital learning platform from the perspective of nine experts is reaching the most appropriate level. Table 4 shows the details of the system evaluation results.

Table 4. System evaluation results from six experts

No.	Assessment Item	Mean Score	*SD*	Interpretation
1.	Visibility of system status	5.00	0.00	Most appropriate
2.	Match between the system and the real world	5.00	0.00	Most appropriate
3.	User control and freedom	5.00	0.00	Most appropriate
4.	Consistency and standards	4.67	0.57	Most appropriate
5.	Error prevention	4.67	0.57	Most appropriate
6.	Memory recall and retrieval system	4.67	0.57	Most appropriate
7.	Flexibility and efficiency of use	5.00	0.00	Most appropriate
8.	Beautiful and simple design	4.33	0.57	Very appropriate
9.	Help users recognize, diagnose, and recover from errors.	4.67	0.57	Most appropriate
10.	The system helps and manuals	5.00	0.00	Most appropriate
11.	Support and enhance user skills	5.00	0.00	Most appropriate
12.	Satisfaction and acceptance of user interactions	4.67	0.57	Most appropriate
13.	Protection of personal information	4.67	0.57	Most appropriate
	Total	4.79	0.40	Most appropriate

D. Results of Pearson Correlation Analysis of between Students' Examination Achievement and their Mathematical Proficiency for Each Dimension in the Automated Open-ended Questions Scoring System

After the researchers identified the effectiveness of the automated open-ended question scoring system of the digital learning platform, researchers would like to further examine whether there is any significant relationship between students' mathematical proficiency levels in each dimension and their examination achievement. The results showed that there were positive, strong, and significant correlations between students' examination achievement in the subject of Measurement and Geometry and their mathematical proficiency level in the mathematical process dimension [r(106) =

0.782, p=0.000] as well as in conceptual structural dimension [r(106) = 0.815, 0.000] at the significant level of 0.01. Table 5 illustrates the details of the intercorrelations findings of the variables.

Table 5. Intercorrelation analysis between Students' Examination Achievement and their Mathematical Proficiency Levels

Subject	Dimension	Mathematical process dimension	Conceptual structural dimension
Measurement and geometry	Mathematical process		0.815**
	Conceptual structural dimension	0.782**	
	Variance	0.000	0.000

DISCUSSION

The primary contribution of this research is the automated open-ended question scoring system has successfully proved its effective and reliability. This is because the results indicated that the trial scores obtained from the automated scoring system were consistent with their actual grades which were checked by their teachers manually. Moreover, the automated open-ended question scoring system also provided formative feedback for both teachers and students to diagnose their mathematical proficiencies because teachers can know directly what students learn, what students need to learn, and how well they understand it from the assessment report section. In conclusion, the automated open-ended question scoring system can be used to assist students' self-regulated learning and instructions based on multiple proficiencies.

ACKNOWLEDGMENT

This work was supported by the National Research Council of Thailand (NRCT) (Research grant number: 2564NRCT322021). The authors gratefully acknowledge the use of service and facilities of the Faculty of Education, Khon Kaen University, Khon Kaen 40002, Thailand.

REFERENCES

[1] Y-M, Huang, S-H, Huang, and T-T Wu, "Embedding diagnostics mechanisms in a digital game for learning mathematics," Education Technology Research and Development, vol.62, pp.187-207, 2014. https://doi.org/10.1007/s11423-013-9315-4

[2] S. Chinjunthuk, P. Junpeng, and K. N. Tang, "Use of digital learning platform in diagnosing seventh grade students' mathematical ability levels," Journal of Education and Learning, vol.11, no.3, 2022. https://doi.org/10.5539/jel.v1n3p95

[3] P. Junpeng, M. Marwiang, S. Chiajunthuk, P. Suwannatrai, K. Chanayota, K. Pongboriboon, K. N. Tang, and M. Wilson, "Validation of a digital tool for diagnosing mathematical proficiency," International Journal of Evaluation and Research in Education, vol.9, no.3, pp.665-674, 2020. https://doi.org/10.11591/ijere.v9i3.20503

[4] E. Sung and R. E. Mayer, "Five facets of social presence in online distance education," Computer in Human Behavior, vol. 28, pp. 1738-1747, 2012.

[5] M. Ouadoud, A. Nejjari, M. Y. Chkouri, and K. E. EI-Kadiri, "Learning management system and the underlying theories," In Proceedings of the Mediterranean Symposium on Smart City Application (pp. 732-744), Springer, 2017.

[6] D. Adom, J. A. Mensah, and D. A. Dake, "Test measurement, and evaluation: Understanding and use of the concepts in education," International Journal of Evaluation and Research in Education, vol.9, no.1, pp.109-119, 2020. https://doi.org/10.11591/ijere.v9i1.20457

[7] M. D. Reckase, Multidimensional item response theory: Statistics for social and

behavioral sciences. New York, NY: Springer, 2009.

[8] T. C. Reeves, "Design research from a technology perspective," In J. V. D. Akker, K. Gravemeijer, S. McKenney, & N. Nieveen (Eds.), Educational Design Research (pp. 52-66), New York, NY: Routledge, 2006.

[9] S. Vongvanich, Design research in education, Bangkok, Thailand: Chulalongkorn University Printing House.

[10] T. George, Mixed methods research – Definition, guide & examples, July 21, 2022. [Online]. Available: Mixed Methods Research | Definition, Guide & Examples (scribbr.com)

[11] A. Hayes, "Stratified random sampling," Investopedia, October 4, 2021. [Online]. Available: https://www.investopedia.com/terms/stratified_random-sampling.asp/

[12] S. Kanjanawasi, New theory of testing (3rd ed.). Bangkok, Thailand: Printing House Chulalongkorn University.