Analysis of STACK Answer Data Using Pen-Stroke Data from a Calculation Notebook and Item Response Theory

https://doi.org/10.3991/ijet.v18i01.36531

Yasuyuki Nakamura(✉), Kazusa Fujimoto, Kako Ito
Nagoya University, Nagoya, Japan
nakamura@nagoya-u.jp

Abstract—One of the most significant features of STACK, an automatic assessment system for mathematics, is that it can automatically classify correct, semi-correct, and incorrect answers using a Potential Response Tree and provide appropriate feedback for each. However, answers provided by students are only the results of their calculations, and if they are semi-correct or wrong, it is necessary to check the notes describing the calculations to determine where mistakes have been made. We developed a function that allows students to submit their notes at the same time as their STACK answers, thereby creating an environment that allows teachers to check where students have made mistakes and where they have stumbled. The note submission can be in the form of saving the results of calculations made with a digital pen on a tablet. This study examines a method of analysing answer data by visualising and analysing pen-stroke data from notes submitted using the latter method and linking this to correct/incorrect data. From pen-stroke data, two types of visualisation were achieved on the submitted notes. One was to colour-code the strokes according to the writing speed of the digital pen, and the other was to display the strokes as thicker if the stagnant time during which nothing was written was above a threshold value. The difficulty of the problem was surmised from those data. Then, item response theory was used to verify whether the difficulty was correctly estimated.

Keywords—automatic assessment system for mathematics, pen-stroke data analysis, item response theory

1 Introduction

In recent years, the development of information technology has accelerated informatisation in the education sector, resulting in more attention being paid to e-learning than ever before, as lectures are delivered online and home study is conducted with on-demand materials. One important function of e-learning is online testing, but the majority of test formats are true/false, multiple-choice, or numerical input. However, for online tests of science and mathematics, when the answer method is multiple-choice, it is possible to guess the answer to some extent from the options, and even if the calculation method is not known, it may be possible to guess the correct answer. For example, in integral questions, it is possible to find the correct answer by differentiating all the functions given as options, even if one does not understand how to calculate the integral. Where answers are multiple-choice, it may be possible to assess overall ability by assigning a large number of questions, but it is difficult to ascertain exactly how well students understand individual questions. On the other hand, in the case of an answer format in which formulae are entered, it is expected that the actual ability of the respondent can be measured as it is impossible to answer a question if the calculation method is not understood. In the case of a formula input format, it is also expected that it is possible to obtain information on areas where students who have taken the online test lacked sufficient understanding by analysing wrong answers.

Automatic formula scoring systems include STACK [1], Möbius [2], WeBWorK [3] and Numbas [4]. This paper examines a method for estimating students' abilities using answer data from STACK, which is the most widely used system in Japan. Specifically, this paper uses the function for submitting notes in which calculations are made to provide answers, which will be introduced in the next section, and features of the answers are extracted from the written data of the notes to estimate the difficulty level of the questions and the students' level of understanding. Furthermore, from the correct and incorrect information from the online test, the ability values of the students who took the test, and the difficulty level of the questions estimated using item response theory are calculated with the aim of confirming the validity of the findings from the data in the notes.

The paper is structured as follows. Section 2 briefly introduces STACK, and Section 3 introduces the methods utilised in this study. Section 4 gives examples of the application of these methods to real answer data. Section 5 provides a summary and discussion.

2 Brief review of STACK

In STACK and other automatic formula scoring systems, a correct or incorrect evaluation is made by entering a formula as a solution, such as in a problem-solving ordinary differential equation, as shown in Figure 1. If the left-hand side of Figure 1 is the correct answer and the integral constant is missing, as in the right-hand side of Figure 1, partial points can be given. This kind of partial point evaluation of answers can be achieved with a function called Potential Response Tree.

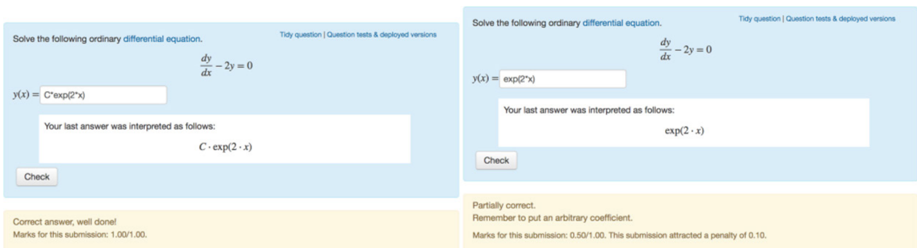


Fig. 1. Example of STACK

The potential response tree is an algorithm that enables the classification of student answers by evaluating the answers step by step and from several perspectives by means of a dichotomous tree. For example, in the potential response tree in Figure 2, first, at node 1, it is determined whether the student's answer satisfies the differential equation or not, and if it is true, it proceeds to node 2; but if it is false, it is scored as 0. At node 2, it is determined whether the student's answer is one that neglects an arbitrary constant, and if true, a partial score of 0.5 points is awarded; but if false, it proceeds to node 3. At node 3, it is determined whether the student's answer is a trivial one, and if true, it is awarded 0.1 partial points; but if it is false, 1 point is awarded because it is a correct answer.

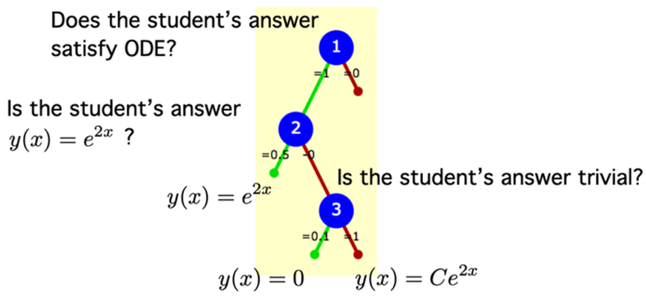
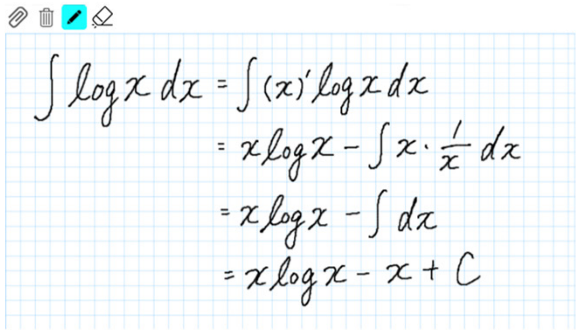


Fig. 2. An example of a potential response tree of STACK for assessing the question about ordinary differential equations

3 Pen-stroke data analysis and item response theory

Nakamura and Nakahara developed a function that allows the submission of notebooks containing not only answers but also notes describing the calculation process, which can be linked to and managed in relation to questions and answers [5]. Notes can be submitted either by attaching handwritten notes with photographs or by writing on a tablet with a digital pen. An example of a notebook written on a tablet with a digital pen is shown in Figure 3 (left). In the upper left-hand corner of the notebook, there are buttons for selecting the end of the note, erase all, pen, and eraser. The digital pen is used to write answers on the grid. The pen-stroke data contains the pen-strokes written by the students in chronological order. Figure 3 (right) shows a part of the recorded data. Action indicates the state of the pen, such as writing start, writing in progress, or selecting the eraser; X and Y are the coordinates of the pen nib; Time indicates the UNIX time when the Action was performed.



Action	X	Y	Time
start-drawing	66	38	1639025496916
move-drawing	66	37	1639025496927
move-drawing	65	33	1639025496938
move-drawing	73	214	1639025502317
changetoeraser	0	0	1639025507711
start-erasing	81	204	1639025508124
move-erasing	80	204	1639025508136
move-erasing	67	207	1639025508609
changetopen	0	0	1639025509592
start-drawing	455	96	1639025510342
move-drawing	455	95	1639025510351
move-drawing	724	531	1639025547555
allclear	0	0	1639025571671
start-drawing	101	66	1639025572372
move-drawing	695	819	1639025693828
savenote	0	0	1639025706243

Fig. 3. An example of a note submitted and a part of the recorded data

The pen stroke data shown in Figure 3 is used to visualise the writing speed and stagnation points. Based on the writing speed obtained, colour is added for each speed. The lower the speed, the bluer the colour; the higher the speed, the redder the colour. If more than 2 seconds have elapsed between the end of a stroke and the start of the next stroke, the post-stagnation stroke is displayed in bold as a stagnation point. Two seconds or more is a length that can be perceived as stagnation by humans [6].

According to previous empirical findings, students who rewrite more often are less certain, but students who write slower but rewrite less tend to solve problems more carefully and accurately. However, this assessment is subjective, and so, an objective evaluation is required. We, therefore, used item response theory to estimate students' ability and question difficulty from the correctness and incorrectness data of the test and compared them with the findings from the pen-stroke data to check the consistency.

4 Analysis of answer data

The visualisation method described in section 3 was applied to the answer data of three students who answered six differential and integral calculus questions. The correct/incorrect data and the number of times the eraser was used by each student for each question are summarised in Table 1. Figure 4 shows the notes of two of the students' answers to Question/Item 3. The number of times the total erasure and eraser were used in this question was 2 for student 1, 0 for student 2 (left of Figure 4), and 4 for student 3 (right of Figure 4). The results for the colour of the letters show that student 2 changed his writing speed frequently. The thickness of the first stroke of the equal sign also changed. Student 3's colour did not change much, indicating that the student's writing speed is almost constant. The thickness of the letters changes in only one place, and the thickness is not much different from normal. The number of times the eraser was used was higher for the question with more incorrect answers than for the question with more correct answers. Student 2 solved the questions after thinking about how to solve them is indicated by the speed changing rapidly and there being some stagnation.

However, student 3 solved the questions after thinking about how to solve them while writing the answers, which is indicated by the speed not changing much and there being little stagnation. It is also possible that the speed and stagnation are related to the way the intermediate equation is written. It has also been assumed that the difficulty of a problem corresponds to the number of times it is eliminated. Therefore, Questions/Items 4, 5, and 6 could be guessed to be relatively difficult. Student 1 is predicted to be competent, as this student correctly answered question/item 6, which is estimated to be more difficult, although the student had to redraw it more often.

Table 1. Summary of correct and incorrect results for each question/item and the number of times the eraser was used in the questions of differentiation, integral, and story

	Differentiation				Integral						Story	
	Item1		Item2		Item3		Item4		Item5		Item6	
	1/0*	e**	1/0	e	1/0	e	1/0	e	1/0	e	1/0	e
Student1	1	0	1	0	1	2	0	5	1	7	1	32
Student2	0	0	0	0	1	0	0	4	0	12	0	14
Student3	1	2	1	0	0	4	0	3	0	8	0	12

Notes: *Results: 1; correct, 0; incorrect, **Number of times erasers are used.

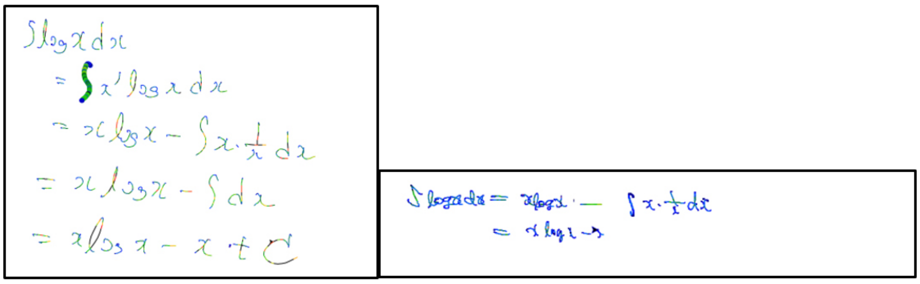


Fig. 4. An example of a note submitted and a part of the recorded data for student 2(left) and student 3(right)

Next, item response theory was used to estimate a student's ability and the difficulty of the problem. The ability values were Student 1 > Student 3 > Student 2, which was roughly inferred in the previous section. Based on the correct/incorrect data, the discriminative ability a and difficulty b of the equation

$$p_i(\theta) = \frac{1}{1+e^{-1.7a_i(\theta-b_i)}} \tag{1}$$

were estimated, where $p_i(\theta)$ is the probability of a student with ability θ answering question i correctly. Equation (1) with calculated a and b, which are called item response curves, are illustrated in Figure 5. This figure indicates that questions/items 4 and 6 are more difficult. The results predicted in previous paragraphs are here demonstrated objectively.

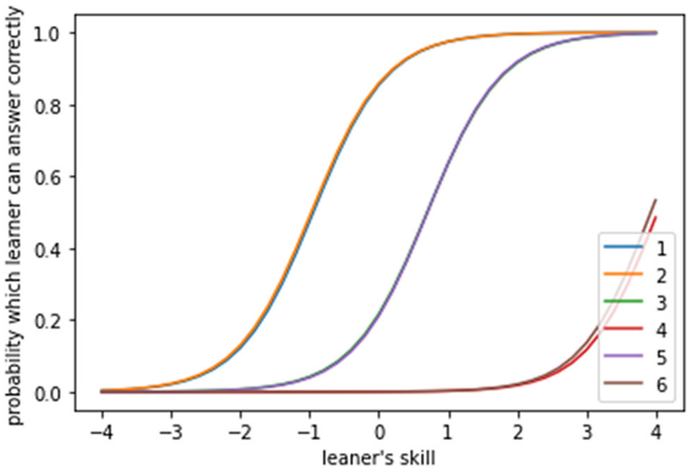


Fig. 5. Item response curves with calculated *a* and *b*

5 Summary

In this paper, which uses STACK, we visualised the solution process, writing speed, and stagnation points based on pen-stroke data obtained from the solution process of mathematical problems written on a tablet and also predicted students' ability and the difficulty level of problems. The results were matched to the results of item response theory, and qualitative results were obtained. However, due to the small number of data, it is difficult to qualify this as a general trend, and it will be necessary to increase the number of data and verify this trend in more detail in the future.

6 Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers 21H04412, 19K21758.

7 References

[1] C. Sangwin, *Computer Aided Assessment of Mathematics*, Oxford University Press, 2013. https://doi.org/10.1093/acprof:oso/9780199660353.001.0001

[2] Möbius – DigitalEd. [Online]. Available: https://www.digitaled.com/mobius/. [Accessed Oct. 31, 2022].

[3] The WeBWorK Project. [Online]. Available: https://openwebwork.org/. [Accessed Oct. 31, 2022].

[4] Numbas. [Online]. Available: https://www.numbas.org.uk/. [Accessed Oct. 31, 2022].

[5] Y. Nakamura and T. Nakahara, "Note-Submission Function for Moodle Quiz and Collecting Pen-Stroke Data," in IADIS International Conference Mobile Learning 2019, 2019, pp. 163–164. https://doi.org/10.33965/ml2019_201903P001

[6] Masaaki Iiyama, Chihiro Nakatsuka, Yoshitaka Morimura, Atsushi Hashimoto, Masayuki Murakami, Michihiko Minoh, "Detecting Answer Stuck Point Using Time Intervals of Pen Strokes", *Jpn. Soc. Infor. & Sys. Edu*, vol. 34, no. 2, pp. 166–171, 2017.

8 Authors

Yasuyuki Nakamura is an associate professor at the Graduate School of Informatics of Nagoya University, Nagoya, 464-8601, Japan.

Kazusa Fujimoto is a graduate student at the Graduate School of Informatics of Nagoya University, Nagoya, 464-8601, Japan.

Kako Ito is a graduate student at the Graduate School of Informatics of Nagoya University, Nagoya, 464-8601, Japan.

Article submitted 2022-10-03. Resubmitted 2022-11-14. Final acceptance 2022-11-18. Final version published as submitted by the authors.