Students' experiences with automated final answer diagnoses for mathematics tasks

Gerben van der Hoek[1], Bastiaan Heeren[2], Rogier Bos[1], Paul Drijvers[1] and Johan Jeuring[1]

[1]Utrecht University, the Netherlands; g.vanderhoek@uu.nl

[2]Open University, the Netherlands

Model Backtracking (MBT) is a novel technique for automated detailed diagnoses based on final answers. In this small-scale pilot study, we answer research questions about nine 15-to-17-year-old senior general secondary students' experiences with MBT diagnoses. The students practiced linear extrapolation in a learning environment that provides error-specific feedback and selects appropriate subtasks using MBT. Data included screen captures of students navigating the environment and interviews on students' experiences with the environment. Results showed approaches ranging from correcting an error after receiving feedback to trial-and-error behavior while repeatedly consulting the worked-out solution. Furthermore, students preferred error-specific feedback over worked-out solutions. They found that worked-out solutions provide an insightful overview; yet, errors are not pinpointed, and worked-out solutions reduced motivation for further practice.

Keywords: Automated assessment, Feedback, Linear extrapolation, Model Backtracking,

Introduction

The use of automated formative assessment in mathematics education is rapidly increasing. Various software tools are available to give students feedback on mathematical calculations and underlying strategies. Many of these tools either provide feedback that is non-specific to the student error or require all calculation steps as input. Inputting every step of a calculation can be cumbersome for a learner (Drijvers, 2019). Moreover, software should allow a student to authentically do mathematics, without unintended effects caused by the environment's interface (Kieran and Drijvers, 2006). To provide error-specific feedback, the interface for calculations that consist of several nonequivalent steps (e.g., linear extrapolation) often contains an input box for each step. However, in this way, a calculation is pre-structured in the interface and a student's reasoning might be unintentionally scaffolded by this structure. This unintended effect can be avoided by using final answer diagnoses through Model Backtracking.

Model Backtracking (MBT) (van der Hoek, 2020; van der Hoek et al., 2023) is a technique that uses the strategy language (Heeren et al., 2010) to provide a detailed diagnosis of an entire student calculation based on a final answer. MBT itself is comprised of several techniques that serve two purposes. The first purpose is to mitigate the combinatorial explosion associated with calculating all possible final answers, given a set of possible student errors. The second purpose is to increase the accuracy of final answer diagnoses by selecting specific tasks. In MBT, the starting parameters for a task are selected such that the number of different ways to reach a final answer is minimized.

MBT allows for providing error-specific feedback based on a final answer. This might help students in postponing viewing worked-out solutions to tasks they should learn from by solving them. Students

tend to alleviate their feelings of uncertainty (Shute, 2008) by viewing worked examples. However, viewing these work examples might prevent students from developing problem-solving skills (Goodman and Wood, 2004). This paper reports on a small-scale qualitative pilot study on how senior general secondary 15 to 17-year-old students experience working with error-specific feedback in our online environment. To investigate these experiences, we formulated two research questions:

1. How did the students use the error-specific feedback and the worked-out solutions provided by the environment?
2. How did the students describe experiences their with error-specific feedback as opposed to worked-out solutions?

In the next section, we identify a theoretical framework suitable for answering these questions.

Theoretical framework

Online learning environments offer functions that are grounded in the following pedagogical approaches: learning from feedback, learning from worked examples, and learning from tasks. Here, we explore these approaches further. We start by listing various feedback types relevant to our research. Shute (2008) presents several feedback types, such as verification, try-again, and elaborated feedback. Below we further elaborate on these types. Verification feedback provides learners with knowledge about the correctness of a response, which is often referred to as knowledge of results (KR). Try-again feedback (TA) allows learners to provide a new response after some other type of feedback is provided. As for elaborated feedback, Shute distinguishes several variants, two of which are of interest here: Topic-contingent feedback and feedback on bugs. The former is feedback about the topic that is being studied, which could be a worked example (WE) of a task; the latter is error-specific feedback (ES), which is based on a diagnosis of the learner's response.

Feedback can benefit learners in two ways: it can resolve uncertainty and it can alleviate cognitive load. We discuss uncertainty and cognitive load in more detail since these are important aspects of providing feedback. Uncertainty is an unpleasant state that may distract learners from task performance; hence, they wish to avoid or resolve it (Bordia et al., 2004). Therefore, providing feedback can increase performance by alleviating this uncertainty. Cognitive load is introduced when task execution floods learners' working memory. Cognitive load can be decreased by using worked examples. For instance, Sweller et al. (1998) showed that worked examples reduced the cognitive load for low-ability students in problem-solving tasks.

Learning from problem-solving activities as opposed to learning from worked examples has been studied in the late eighties and nineties (Chi et al., 1989, Renkl 1997). These studies generally favored learning from worked examples; because the cognitive load that problem-solving activities introduce can cloud the actual learning process. However, Chi et al. also reported that positive learning outcomes using worked examples strongly depend on a student's ability to self-explain the steps in the worked example. Furthermore, Goodman and Wood (2004) found that very specific feedback can be detrimental to long-term performance, as it may prevent learners from developing problem-solving skills.

In conclusion: when learners perform a task, they can experience an adverse state of cognitive load and uncertainty. As such, they wish to alleviate this state by seeking feedback. Therefore, feedback is a useful tool to help the learning process. Nonetheless, caution is warranted for its use; because worked examples only benefit students who possess self-explaining skills, and feedback that is too specific can hinder self-development.

Methods

The methods section consists of two parts. We first elaborate on the design of our online environment. After that, we proceed to explain the methods used to conduct the experiment involving the students.

Environment design

The environment consists of a website together with an MBT script that calculates feedback or suggests a subtask based on a student's input. The environment offers a main task and three subtasks. All tasks have random starting values based on 50 pre-calculated integer task parameters that ensure high diagnosis accuracy. This allows a student to retry a task with different starting values after KR, ES, or WE feedback.

Main task format:

Given the table with values for x_1, x_2, y_1, y_2 and x_v

x	x_1	x_2	x_v
y	y_1	y_2	y_v

Q: Use linear extrapolation to compute y_v.

Subtask A: Simpler numbers.

Given the table with values for x_1, y_1, y_2 and x_v

x	x_1	$x_1 + 1$	x_v
y	y_1	y_2	y_v

Q: Use linear extrapolation to compute y_v, first determine the change in y when x increases with 1.

Subtask B: Known slope.

Given the table with values for x_1, y_1 and x_v and known slope a

x	x_1	x_v
y	y_1	y_v

Q: Suppose the slope is a, use linear extrapolation to compute y_v.

Subtask C: Calculate slope.

Given the table with values for x_1, x_2, y_1 and y_2

x	x_1	x_2
y	y_1	y_2

Q: Calculate the slope.

The feedback in the environment is designed following findings in the literature. The system provides KR feedback, and when a diagnosis of a student error is possible it provides ES feedback. The student then has the option to try again to obtain TA feedback. The ES feedback in the main tasks has low specificity (i.e., verbally formulated suggestions), whereas the ES feedback in the subtasks has higher specificity (i.e., suggestions that may contain calculations). WE feedback, a worked-out solution, is available; but we postpone the worked-out solution until after a student has selected a subtask. A student, however, is at liberty to immediately select a subtask and view the worked-out solution. Once a student returns to the main task WE feedback will be available for the main task.

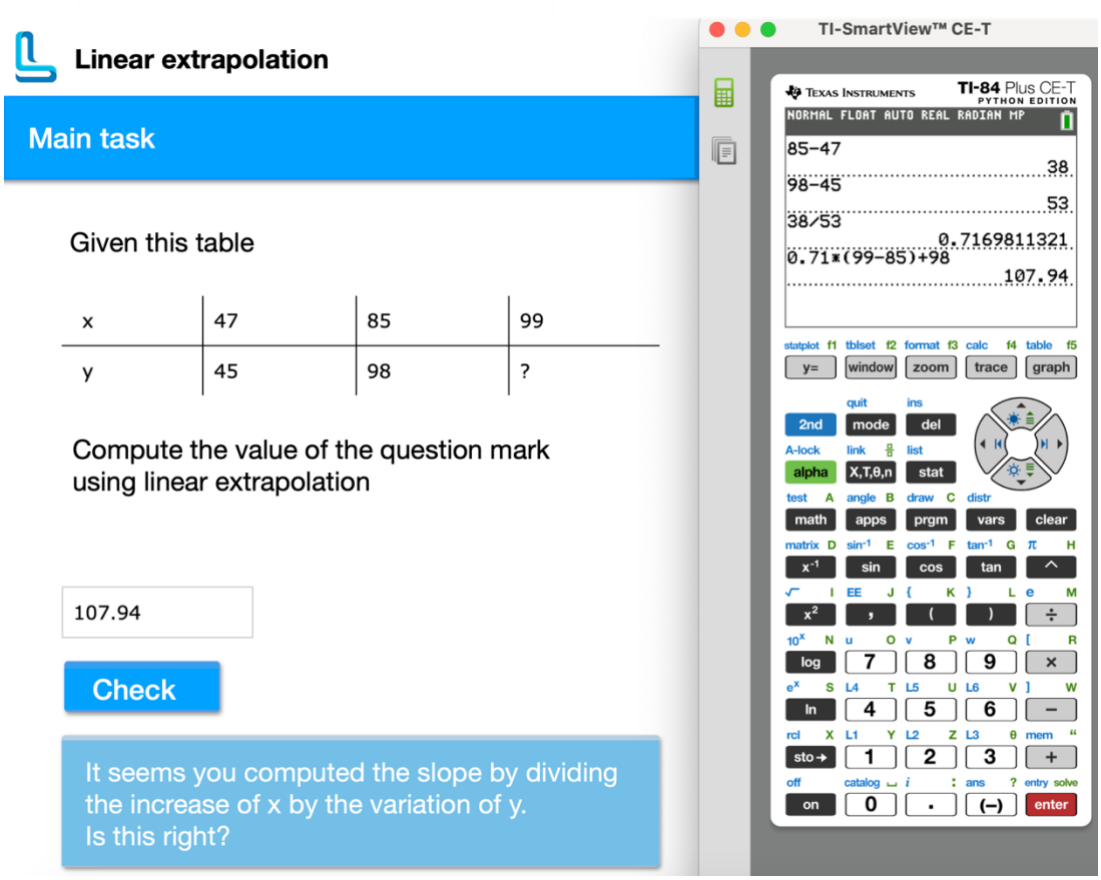

Figure 1: Example of ES feedback during a task

Figure 1 shows ES feedback a student receives when inversely computing the slope. To detect the various (combinations of) student errors, so-called buggy rules are implemented in the system. These buggy rules represent erroneous steps in a student's calculation. Over 24 rules are implemented. The rules are based on empirical findings by Van der Hoek et al. (2023) and work by Van Dooren et al.

(2005) on the unwarranted use of proportional models in missing value problems. The rules include, for instance, using a proportional model (i.e., calculating: $y_v = y_1/x_1 \cdot x_v$), inversion of the slope (i.e., using: $\Delta x/\Delta y$), and (in)correct intermediate rounding of the slope. Subsets of the buggy rules for the main task were used for the subtasks.

In the environment, a student has several options presented by buttons: (1) go to a subtask that is selected by the system based on a diagnosis, (2) retry the current task with different starting values, and (3) view a worked-out solution (when available). The subtask is selected by the MBT system, through the following arrangement: Subtask A is selected in case no error could be detected, a student has provided no input yet, or a student calculates: $y_v = y_2 + (y_2 - y_1)$. Subtask B is selected when a student correctly calculated the slope (rounding errors are allowed) but made a detectable error elsewhere. Subtask C is selected when a student detectably calculates the slope incorrectly.

To view how a student navigated the environment, see https://youtu.be/YYRkew5-EEI for an example replayed at 10 times the normal speed. In the next subsection, we further explain how the experiment in which students use the environment was set up.

Data and data analysis

For this qualitative experiment, a convenient sample of eight senior general secondary students from 10th grade and one student from 11th grade were recruited from four different classes in the school in the Netherlands where the first author is employed. Participation was based on availability and consent to partake. The students had received prior education on linear extrapolation as part of their standard curriculum, but not within four weeks before the experiment.

The students were invited to complete the main task in the environment in a session ranging between 10 and 30 minutes depending on the student. Screen captures along with audio recordings were used to document the students using the environment. Pen, paper, and an onscreen graphic calculator were available to the students. A researcher supported the students in case of confusion on how to operate the system, but not in case of confusion on the task. After the session with the environment ended, the researcher conducted a semi-scripted interview to determine the experiences of the students with the environment.

The screen captures of the sessions with the environment were transcribed into chronological accounts of the events. These accounts were then ordered according to the amount of help the student required from the system. From this ordering five different approaches to using the error-specific feedback and the worked-out solutions emerged. General descriptions of the approaches are formulated and summarized in Table 1. The students' utterances in the interviews that showed the experiences of the students with worked-out solutions and error-specific feedback were transcribed. These utterances were grouped by similarity and coded with a common theme.

Results

In Table 1 we find descriptions of students' approaches to using the error-specific feedback and the worked-out solutions. The *help level* ranges from requiring no help at all from the system (0) to requiring much help (4). KR feedback provides knowledge of results, ES feedback is error-specific

feedback, WE feedback consists of a worked example and TA feedback allows for retries after KR, WE, or ES feedback.

Table 1: Various usages of help in the environment

Help level	Description of help usage	Frequency
0	The student correctly completed the main task and received KR feedback	1
1	The student corrected an error after ES feedback	2
2	The student solved the tasks by using ES feedback, only using WE feedback when ES feedback was unclear	2
3	The student solved the tasks by studying WE feedback and correcting errors with ES feedback	3
4	The student alternated between TA feedback and WE feedback	1

Using Table 1 we can divide students into two groups, Group A with a help level of 2 or less, and Group B with a help level of 3 or more. The difference between these groups is the use of worked-out solutions. Students in Group A seldom used a worked-out solution and if they did it was because feedback from the system was unclear to them. In contrast, students in Group B used a worked-out solution as a worked example in the sense of Chi et al. (1989). Of these students, three students immediately viewed the worked-out solution of a subtask without trying the subtask. Furthermore, in Group B we also found the student with help level 4, this student exhibited a trial-and-error-like behavior almost frantically switching between retrying a task and viewing the worked-out solution. The difference between group A and group B can perhaps be explained by the level of uncertainty (Bordia et al., 2004) the students experienced that could have been caused by a lack of sufficient prior knowledge.

Aside from students' approaches to working in the environment, we also investigated students' experiences with worked-out solutions and error-specific feedback through interviews. Table 2 summarizes the utterances during the interviews that were conducted after interacting with the environment.

Table 2: Utterances in the interviews

Utterance	Frequency
Error-specific feedback helps to identify the error	7
An error is not pinpointed in a worked-out solution	4
A worked-out solution reduces motivation for further practice	3
A worked-out solution provides an insightful overview of the task	3

For worked-out solutions, we have two negative utterances with a total frequency of 7 (the second and the third from the top) and we have one positive utterance with a frequency of 3 (the fourth). For error-specific feedback, we have only a positive utterance with frequency 7 (the first). If one were to

summarize these results, one could conclude that in this group error-specific feedback is preferred over worked-out solutions.

Conclusion

How did our sample of senior general secondary students experience practicing linear extrapolation in our MBT-driven learning environment? First, we consider the research question on the use of error-specific feedback and worked-out solutions. We found that some students studied the worked-out solutions instead of only using worked-out solutions to check their answers. Studying worked-out solutions can lead to memorizing them without proper self-explanation (Chi et al., 1989). This effect can be reduced by offering error-specific feedback and postponing worked-out solutions.

Three of the nine students in our sample immediately viewed the worked-out solution of a subtask without trying the subtask. Perhaps this can be explained by the students' uncertainty, caused by a lack of sufficient prior knowledge of the tasks. If so, this can be remedied by providing the opportunity for direct instruction, possibly by incorporating an instruction video in the environment. A combination of direct instruction and inquiry is an effective way of learning (De Jong et al., 2023).

Next, we consider the research question on students' experiences with ES (error-specific) feedback as opposed to WE feedback (worked-out solutions). Overall, students had a positive experience with the ES feedback provided by the environment. They found it helped them pinpoint errors in their calculations whereas WE feedback did not. Furthermore, they found that WE feedback provided a clear overview of the task. However, WE feedback reduced motivation for further practice with similar tasks, since there is not much left to explore after studying the worked-out solution. This provides an argument for offering ES feedback and postponing WE feedback.

Now we reflect on the validity of any claims we made. We have a very small sample from a specific group of students, which means that we cannot generalize beyond statements about the existence of certain behaviors or opinions of students. Even the explanations for the various phenomena offered in this section and the previous sections are at best hypotheses. Since these explanations are given *after* the phenomenon was observed, they hold little to no bearing on any claims of cause and effect. Then, what have we gained by this endeavor? We have gained two things, firstly we have gained *leads* for further study and secondly, we have gained *leads* for improvement of the environment.

To summarize, the experiences of students with the MBT environment were overall positive and eight out of the nine students completed the main task using the information provided by the system. Further development of such MBT-driven systems might contribute to improving learning processes. This future research could show that senior general secondary students, with low self-explanation skills, can benefit from practicing procedures such as extrapolation with the aid of error-specific feedback provided by MBT.

References

Bordia, P., Hobman, E., Jones, E., Gallois, C., & Callan, V. J. (2004). Uncertainty during organizational change: Types, consequences, and management strategies. *Journal of Business and Psychology, 18,* 507–532.

Chi, M. T., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive science, 13(2)*, 145–182.

De Jong, T., Lazonder, A. W., Chinn, C. A., Fischer, F., Gobert, J., Hmelo-Silver, C. E., ... & Zacharia, Z. C. (2023). Let's talk evidence–The case for combining inquiry-based and direct instruction. *Educational Research Review*, 100536. https://doi.org/10.1016/j.edurev.2023.100536

Drijvers, P. (2019). Digital assessment of mathematics: Opportunities, issues and criteria. *Mesure et Évaluation en Éducation, 41*(1), 41–66. https://doi.org/10.7202/1055896ar

Fong, C. J., Patall, E. A., Vasquez, A. C., & Stautberg, S. (2019). A meta-analysis of negative feedback on intrinsic motivation. *Educational Psychology Review, 31*, 121–162.

Goodman, J. S., & Wood, R. E. (2004). Feedback specificity, learning opportunities, and learning. *Journal of Applied Psychology, 89*(5), 809.

Heeren, B., Jeuring, J., & Gerdes, A. (2010). Specifying rewrite strategies for interactive exercises. *Mathematics in Computer Science, 3*(3), 349–370. https://doi.org/10.1007/s11786-010-0027-4

Kieran, C., & Drijvers, P. (2006). The co-emergence of machine techniques, paper-and-pencil techniques, and theoretical reflection: A study of cas use in secondary school algebra. International Journal of Computers for Mathematical Learning, 11(2), 205–251. https://doi.org/10.1007/s10758-006-0006-7

Renkl, A. (1997). Learning from worked-out examples: A study on individual differences. *Cognitive science, 21*(1), 1–29.

Sweller, J., Van Merriënboer, J., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251–296.

Tall, D. (2013). *How Humans Learn to Think Mathematically: Exploring the Three Worlds of Mathematics* (Learning in Doing: Social, Cognitive and Computational Perspectives). Cambridge University Press. https://doi.org/10.1017/CBO9781139565202

Van Dooren, W., De Bock, D., Hessels, A., Janssens, D., & Verschaffel, L. (2005). Not everything is proportional: Effects of age and problem type on propensities for overgeneralization. *Cognition and Instruction, 23*(1), 57–86. https://doi.org/10.1207/s1532690xci2301_3

Van der Hoek, G., (2022). Evaluating digital student work through model backtracking. In J. Hodgen, E. Geraniou, G. Bolondi, & F. Ferretti (Eds.), *Proceedings of the Twelfth Congress of the European Society for Research in Mathematics Education (CERME12)* (pp. 2873–2880). Free University of Bozen-Bolzano and ERME.

Van der Hoek, G., Heeren, B., Bos, R., Drijvers, P., & Jeuring, J. (2023). *Model backtracking: Designing and testing an approach to automated diagnosis of students' work through final answers* [Manuscript submitted for publication].